Instrumental Variables and Regression Discontinuity

Fernando Hoces la Guardia
08/01/2022

Housekeeping

- Midterm 2 Grades Wednesday.
- Midterm 2 Solutions: Today.
- Practice questions for new material: collection of reading comprehension questions at the end of each chapter (will post IV today).

Combining IV and Regression: 2SlS

- Two reasons to combine IV with regression:

  1. Sometimes we might have more than one instrument and combining them in one regression improves statistical precision (because of a smaller variance in the residual).

  2. Our instruments might not be "as-good-as-random" but might achieve independence after controlling for a few observable characteristics (e.g. age of the mother in case of the twins instrument).

- The procedure that combines regression and IV is called **Two Stage Least Squares (2SLS)**

First Stage and Reduce Form in Regression

- For the case of a binary instrument, we can write the first stage and reduce form as the following regression (end of lecture on CEF):

$$\begin{aligned} \textbf{THE FIRST STAGE:} \quad & D_i = \alpha_1 + \phi Z_i + e_{1i} \\ \textbf{THE REDUCED FORM:} \quad & Y_i = \alpha_0 + \rho Z_i + e_{0i} \end{aligned}$$

- Where we can evaluate each conditional expectation from the previous formulation (of FS and RF) and obtain:

$$\begin{aligned} \textbf{THE FIRST STAGE:} \quad & E[D_i|Z_i = 1] - E[D_i|Z_i = 0] = \phi \\ \textbf{THE REDUCED FORM:} \quad & E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0] = \rho \end{aligned}$$

- Where $LATE = \lambda$ is the ratio the slopes of both regressions.
- 2SLS offers an alternative way of computing this ratio (and getting the SEs right!)

2SLS Procedure

- First step: estimate the regression equation for the first stage and generate fitted values $\widehat{D}_i$:

$$\widehat{D}_i = \alpha_1 + \phi Z_i$$

- Second step: regress Y_i on $\widehat{D}_i$:

$$Y_i = \alpha_2 + \lambda_{2SLS}\widehat{D}_i + e_{2i}$$

- The regression estimate for λ_{2SLS} is **identical** to the ratio ρ/ϕ! (proved in the appendix of Ch3)

2SLS With Multiple Regressors

- Now that we have the regression setup ready, it is straight forward to add control.
- The most important thing to remember is that you need to include the additional controls in all the equations (otherwise we would be inducing a type of OVB).
- Using the example of the additional control of maternal age, A_i:

$$\begin{aligned} \textbf{THE FIRST STAGE:} \quad & D_i = \alpha_1 + \phi Z_i + \gamma_1 A_i + e_{1i} \\ \textbf{THE REDUCED FORM:} \quad & Y_i = \alpha_0 + \rho Z_i + \gamma_0 A_i + e_{0i} \end{aligned}$$

And in the 2SLS estimate:

$$\begin{aligned} \textbf{FIRST STAGE FITS:} \quad & \widehat{D}_i = \alpha_1 + \phi Z_i + \gamma_1 A_i \\ \textbf{SECOND STAGE:} \quad & Y_i = \alpha_2 + \lambda_{2SLS} \widehat{D}_i + \gamma_2 A_i + e_{2i} \end{aligned}$$

- 2SLS gets the SEs right for λ_{2SLS} (more on appendix of Ch3).

2SLS With Multiple Instruments

- In addition the twins instrument (Z_i), we can add now the siblings gender instrument. Let's label this last one W_i to avoid confusions. We can also bring the additional controls (Age, A_i, First born boy B_i) and get new first stage:

$$\begin{aligned} \textbf{FIRST STAGE:} \quad & D_i = \alpha_1 + \phi_t Z_i + \phi_s W_i + \gamma_1 A_i + \delta_1 B_i + e_{1i} \\ \textbf{REDUCED FORM:} \quad & Y_i = \alpha_0 + \rho_t Z_i + \rho_s W_i + \gamma_0 A_i + \delta_0 B_i + e_{0i} \end{aligned}$$

- And the corresponding 2SLS estimation:

$$\begin{aligned} \textbf{FIRST STAGE FITS:} \quad & \widehat{D}_i = \alpha_1 + \phi_t Z_i + \phi_s W_i + \gamma_1 A_i + \delta_1 B_i \\ \textbf{SECOND STAGE:} \quad & Y_i = \alpha_2 + \lambda_{2SLS} \widehat{D}_i + \gamma_2 A_i + \delta_2 B_i + e_{2i} \end{aligned}$$

- Ready to read results from most IV papers!

IV Results for Family Size and Education: First Stage

TABLE 3.4
Quantity-quality first stages

	Twins instruments		Same-sex instruments		Twins and same-sex instruments
	(1)	(2)	(3)	(4)	(5)
Second-born twins	.320	.437			.449
	(.052)	(.050)			(.050)
Same-sex sibships			.079	.073	.076
			(.012)	(.010)	(.010)
Male		−.018		−.020	−.020
		(.010)		(.010)	(.010)
Controls	No	Yes	No	Yes	Yes

Notes: This table reports coefficients from a regression of the number of children on instruments and covariates. The sample size is 89,445. Standard errors are reported in parentheses.

IV Results for Family Size and Education: Second Stage + OLS

TABLE 3.5
OLS and 2SLS estimates of the quantity-quality trade-off

		2SLS estimates		
Dependent variable	OLS estimates (1)	Twins instruments (2)	Same-sex instruments (3)	Twins and same-sex instruments (4)
Years of schooling	−.145	.174	.318	.237
	(.005)	(.166)	(.210)	(.128)
High school graduate	−.029	.030	.001	.017
	(.001)	(.028)	(.033)	(.021)
Some college (for age ≥ 24)	−.023	.017	.078	.048
	(.001)	(.052)	(.054)	(.037)
College graduate (for age ≥ 24)	−.015	−.021	.125	.052
	(.001)	(.045)	(.053)	(.032)

Notes: This table reports OLS and 2SLS estimates of the effect of family size on schooling. OLS estimates appear in column (1). Columns (2), (3), and (4) show 2SLS estimates constructed using the instruments indicated in column headings. Sample sizes are 89,445 for rows (1) and (2); 50,561 for row (3); and 50,535 for row (4). Standard errors are reported in parentheses.

IV Results for Family Size and Education: Second Stage + OLS

OLS and 2SLS estimates of the quantity-quality trade-off

Dependent variable	OLS estimates (1)	2SLS estimates: Twins instruments (2)	Same-sex instruments (3)	Twins and same-sex instruments (4)
Years of schooling	−.145	.174	.318	.237
	(.005)	(.166)	(.210)	(.128)
High school graduate	−.029	.030	.001	.017
	(.001)	(.028)	(.033)	(.021)
Some college (for age ≥ 24)	−.023	.017	.078	.048
	(.001)	(.052)	(.054)	(.037)
College graduate (for age ≥ 24)	−.015	−.021	.125	.052
	(.001)	(.045)	(.053)	(.032)

Notes: This table reports OLS and 2SLS estimates of the effect of family size on

IV - Final Considerations 1/2

- Quick intuitions why SE of λ_{2SLS} are wrong if estimated with OLS: $\widehat{D}_i$ is an estimated variable that has more uncertainty that D_i, we know that, but the software doesn't. Hence it generates fictitiously small SEs (SE from 2SLS > SE from OLS).
- When assessing the relevance of one instrument use t-test as usual. When assessing the relevance of multiple (K) instruments use a joint hypothesis test $\phi_1 = \phi_2 = \phi_K = 0$. The rule of thumb here is that the F-statistic reported for these type of tests has to be greater than 10 (p-hacking alert!).
- Beware of studies that are *instrument driven* ("I just found a new cool and clever instrument! Now, which policy could I use this instrument for?") as oppose to *policy driven* ("Policy X is of high relvance, let's look for IVs to identify its causal effect").

IV - Final Considerations 2/2

- When it comes to external validity never forget that LATE is the effect on compliers (MM constantly does!).
- There is a twitter account that emphasizes this extrapolation problem in bio-medical sciences by adding the proper caveat at the end of each new flashy result:

Regression Discontinuity Design

Regression Discontinuity Design

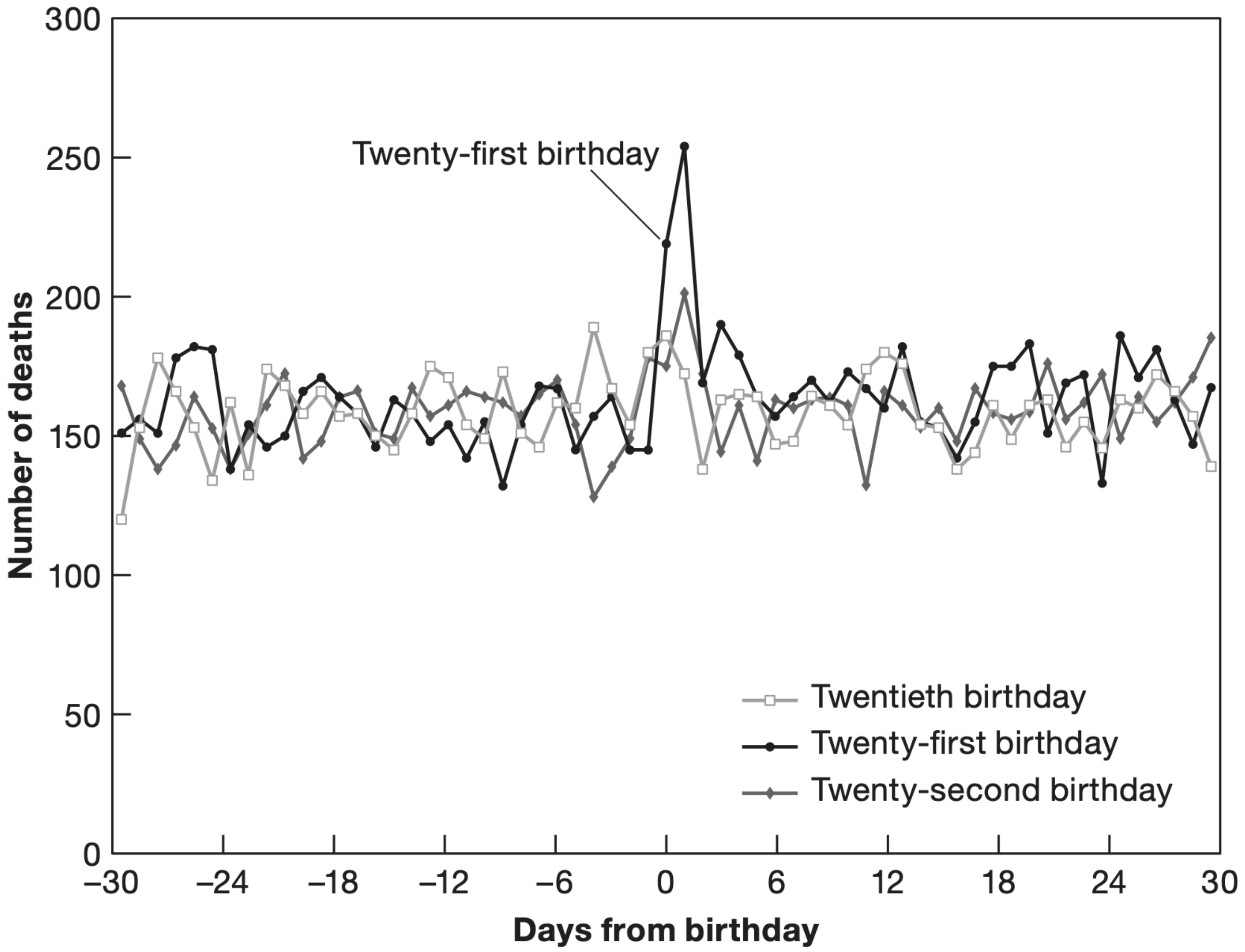
- Many policy decisions (interventions) are assign over the basis of strict rules. For example:
  - California limits the elementary class size at 32.
  - The US federal pensions system (Social Security) starts providing pensions no earlier than at age 62.
  - In order to qualify for certain government programs (e.g. Medicaid in California) families must have an income below a specific threshold.
- Even though these rules seem strict and the opposite of random assignment, we can use them with our fourth research design tool, **Regression Discontinuity Design**, to identify causal effects.

Example: Minimum Legal Drinking Age in the US

- Minimum legal drinking age (MLDA) in the US is 21. Is it too high (or too low)?
  - Advocates: of the current age limit of 21 years old: in some extend reduces access to alcohol, hence preventing harm.
  - Opponents: reducing the drinking age to 18 could discourage binge drinking and promotes a culture of mature alcohol consumption.

Deaths and Distance from Birthdays

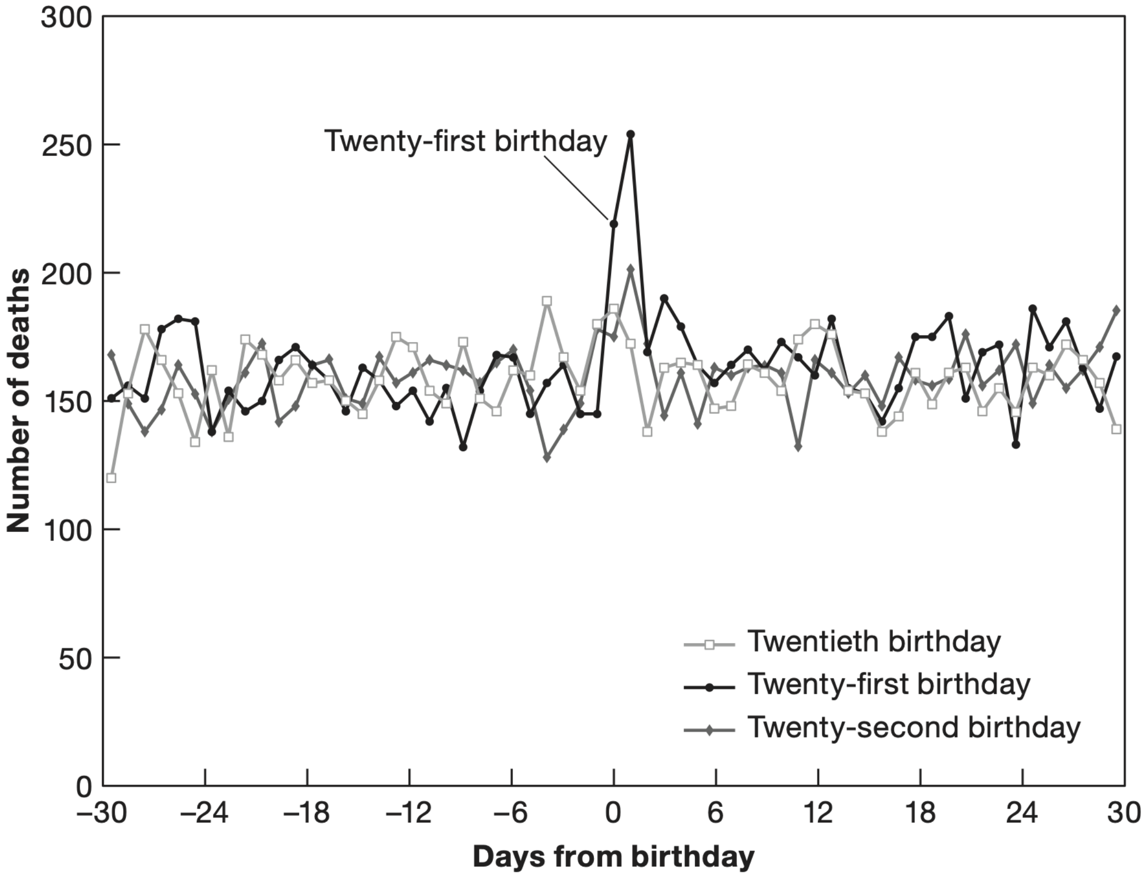
Figure 4.1
Birthdays and funerals



Deaths and Distance from Birthdays. Notes

- Figure shows number of deaths among Americans ages 20-22 between 1997 and 2003. Plotted by day relative to the birthdays. So if somebody was born on January 1st 1990, and died on January 4th 2021, is counted among the deaths of the 21 year old on day 3.
- We will explore this potential effect using RDD
- Spike of about 100 additional deaths per day on the day following the 21st birthday. Over a baseline of 150 deaths (before the spike)
- Nothing similar around other close birthdays (20th or 22nd). We still need to argue that this age-21 effect can be attributed to the Minimum legal drinking age (MLDA) and that it lasts long enough to be worth worrying about.

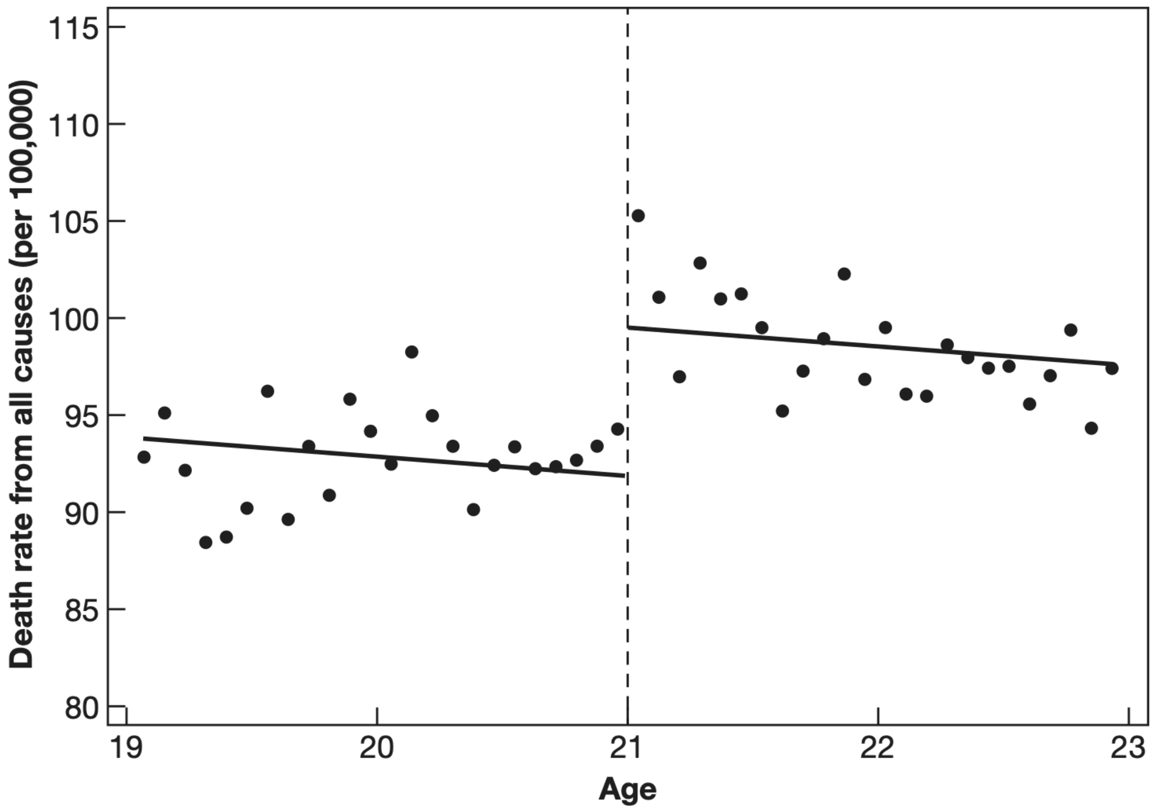
Figure 4.1
Birthdays and funerals



First Exploration of RDD

- Our RDD analysis will focus on these data:
  - Average monthly death rates
  - Months are defined as 30-day intervals, centered around the 21st birthday.
- There is monthly variation but rarely going over 95 deaths per month before the 21st birthday.
- After the 21st birthday, there seems to be an upward shift.
- Also, looking at trends before and after the shift, death rates seem to be decreasing with age. Extrapolating, we should expect deaths (without intervention, or Y_{0i}) to be around 92 (per 100,000) right after the 21st birthdays. They jump instead to around 100.

FIGURE 4.2
A sharp RD estimate of MLDA mortality effects



Notes: This figure plots death rates from all causes against age in months. The lines in the figure show fitted values from a regression of death rates on an over-21 dummy and age in months (the vertical dashed line indicates the minimum legal drinking age (MLDA) cutoff).

RDD Definitions

- **Treatment variable** is D_a, where 1 indicates crossing the legal drinking age (21) and 0 otherwise.
    - Treatment status is a deterministic function of age (a)
    - Treatment status is a discontinuous function of a.
- The variable that determines treatment in RDD, age in this case, is called the **running variable**.
- In a **Sharp RDD** there is a clean switch from control to treatment after crossing a threshold, nobody under the cutoff gets the treatment, and everybody after the

The Regression part of RDD

- The outcome of average mortality for month of age a ($\overline{M}_a$) changes with the running variable for reasons that have nothing to do with the treatment.
- One way to control for this smooth relationship is to add it as a control in a regression like the following:

$$\overline{M}_a = \alpha + \rho D_a + \gamma a + e_a$$

- Estimate of $\rho = 7.7$. Relative to baseline death rate of 95 (without the intervention)
- Is there OVB here?

The Regression part of RDD

- The outcome of average mortality for month of age a ($\overline{M}_a$) changes with the running variable for reasons that have nothing to do with the treatment.
- One way to control for this smooth relationship is to add it as a control in a regression like the following:

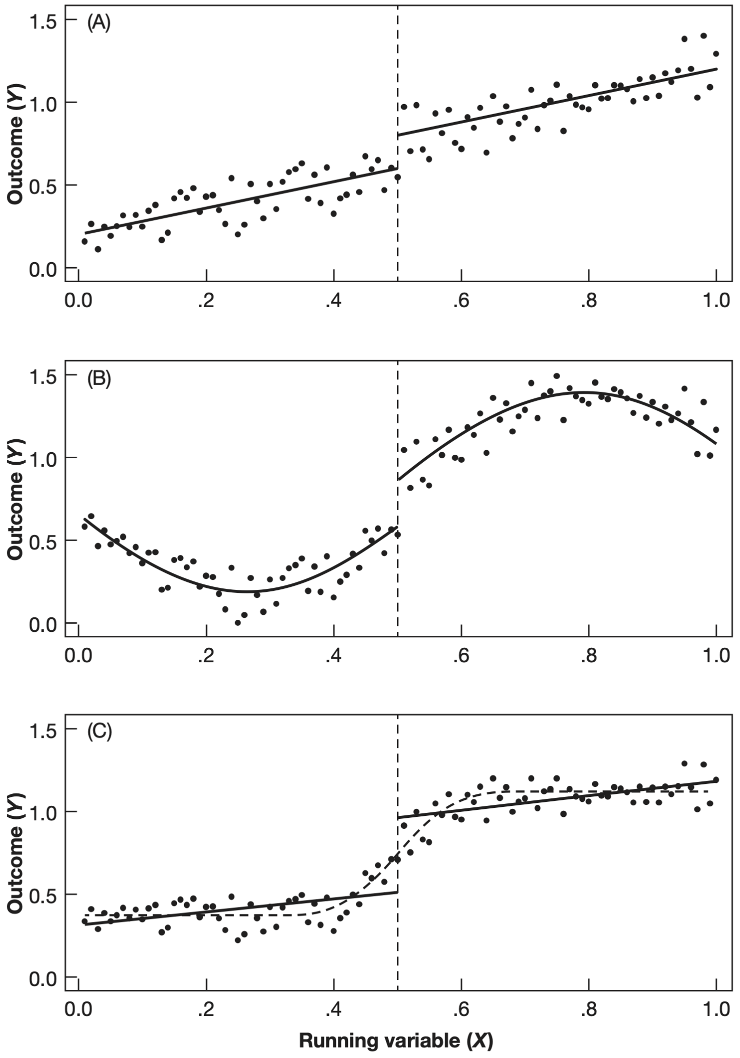
$$\overline{M}_a = \alpha + \rho D_a + \gamma a + e_a$$

- Estimate of $\rho = 7.7$. Relative to baseline death rate of 95 (without the intervention)
- Is there OVB here?
- Given that treatment is a deterministic function of the running variable we know that there is nothing else that affects treatment (so $\pi_1 = 0$ in the auxiliary OVB

But Is There a Jump?

- The key question to identify causality, is whether relationship between running variable and outcome is well represented by a linear control on age.
- Two approaches to reduce the likelihood of mistakes when modeling this relationship: (i) modeling non-linear relationships, and (ii) focusing only on data around the cut-off. We will spend most of the time in (i).
- In addition to logs, non-linearities can be modeled with two additional tools: polynomials and interactions.

Figure 4.3
RD in action, three ways



Notes: Panel A shows RD with a linear model for $E[Y_i|X_i]$; panel B adds

Modeling Non-Linear Relationships: Polynomials

- Curves are usually modeled using polynomials (powers of the regressors).
- Higher polynomials (higher powers) introduce more flexibility but they are also likely to hide a disconitinuity when there is one.
- The choice of how much more flexibility is enough is a judgment call.
- Ideally the results should not vary much as you add higher order polynomials (powers of 3, 4 or more).
- In our example there might be a small curvature in the data, so we add a quadratic term for the running variable:

$$\overline{M}_a = \alpha + \rho D_a + \gamma_1 a + \gamma_2 a^2 + e_a$$

- We are not interested interpreting the effect of age, only on controlling for any non-linear behaviour.

Modeling Non-Linear Relationships: Interactions 1/3

- An interaction is defined as the multiplication of two regressors. Where typically one is a binary regressor.

- Adding an interaction in any regression (or any equation) is a way of capturing changes in (regression) coefficients change for certain groups.

    - Example with just a constant
    - Example with constant and slope
    - Example with both.

- In here we add an interaction and standardize the running variable, so rho can continue to be interpreted as the difference of average outcomes at the cutoff.

Modeling Non-Linear Relationships: Interactions 2/3

- The standardization part might add some confusion, so first let's focus only on adding the interaction to capture a potential shift in the slope that connects age (a) with mortality rates ($\overline{M}_a$):

$$\overline{M}_a = \alpha + \rho D_a + \gamma a + \delta a \times D_a + e_a$$

- The goal of the standardization is to have an easy interpretation of ρ as the difference of mortality around the cut-off. We could define the a new variable $\tilde{a} = a - 21$ which would represent the standardized age $(a - 21)$. This would give us the regression:

$$\overline{M}_a = \alpha + \rho D_a + \gamma \tilde{a} + \delta \tilde{a} \times D_a + e_a$$

Modeling Non-Linear Relationships: Interactions 3/3

- A more generic version would allow for the cut-off to be any number so instead of 21, put a_0. Giving us the standardized formulation of the book:

$$\overline{M}_a = \alpha + \rho D_a + \gamma(a - a_0) + \delta(a - a_0) \times D_a + e_a$$

- The most important part here is understanding the interactions, if you find the standardization distracting, focus on the first two equations but make sure to remember that "we standardize to be able to interpret ρ as the treatment effect"

- (If we want to extrapolate effects awway from the cut-off, we need to be aware that the treatment effect is $\rho + \delta(a - a_0)$)

Non-Linear Relationships: Interactions And Polynomials

Here are polynomials:

$$\overline{M}_a = \alpha + \rho D_a + \gamma_1 a + \gamma_2 a^2 + e_a$$

Non-Linear Relationships: Interactions And Polynomials

Here are polynomials:

$$\overline{M}_a = \alpha + \rho D_a + \gamma_1 a + \gamma_2 a^2 + e_a$$

Here are interactions:

$$\overline{M}_a = \alpha + \rho D_a + \gamma(a - a_0) + \delta(a - a_0) \times D_a + e_a$$

Non-Linear Relationships: Interactions And Polynomials

Here are polynomials:

$$\overline{M}_a = \alpha + \rho D_a + \gamma_1 a + \gamma_2 a^2 + e_a$$

Here are interactions:

$$\overline{M}_a = \alpha + \rho D_a + \gamma(a - a_0) + \delta(a - a_0) \times D_a + e_a$$

Here are combined:

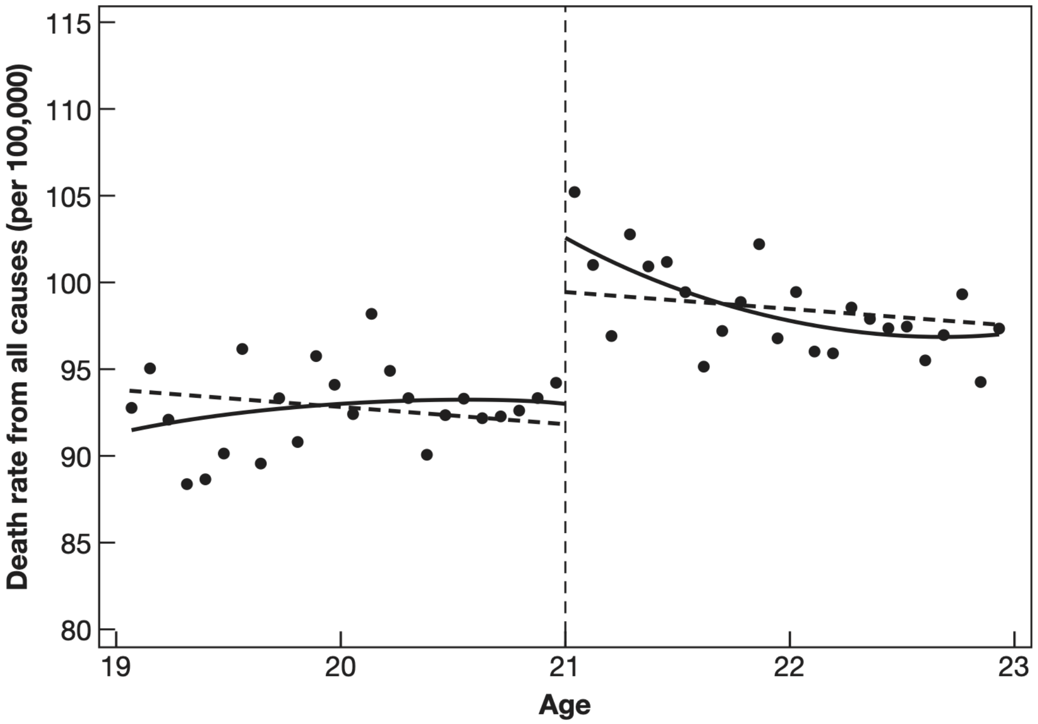
$$\overline{M}_a = \alpha + \rho D_a + \gamma_1(a - a_0) + \gamma_2(a - a_0)^2 + \\ \delta_1 \left[(a - a_0)D_a\right] + \delta_2 \left[(a - a_0)^2 D_a\right] + e_a$$

We can now capture curvature and changing slopes in the relationship between a and $(\overline{M}_a)$, reducing the risk that we incorrectly find a discontinuity where there is none (figure 4.3-C).

The Result

- Effect of 21st birthday seems robust to this new specifications.
- Effect also persist substantially up to the 23rd birthday suggesting lasting effects.
- This last point demonstrate the value of a visual inspection of RDD estimates.

FIGURE 4.4
Quadratic control in an RD design



Notes: This figure plots death rates from all causes against age in months. Dashed lines in the figure show fitted values from a regression of death rates on an over-21 dummy and age in months. The solid lines plot fitted values from a regression of mortality on an over-21 dummy and a quadratic in age, interacted with the over-21 dummy (the vertical dashed line indicates the minimum legal drinking age [MLDA] cutoff).

Now All in One Table

TABLE 4.1
Sharp RD estimates of MLDA effects on mortality

Dependent variable	Ages 19–22 (1)	Ages 19–22 (2)	Ages 20–21 (3)	Ages 20–21 (4)
All deaths	7.66	9.55	9.75	9.61
	(1.51)	(1.83)	(2.06)	(2.29)
Motor vehicle accidents	4.53	4.66	4.76	5.89
	(.72)	(1.09)	(1.08)	(1.33)
Suicide	1.79	1.81	1.72	1.30
	(.50)	(.78)	(.73)	(1.14)
Homicide	.10	.20	.16	−.45
	(.45)	(.50)	(.59)	(.93)
Other external causes	.84	1.80	1.41	1.63
	(.42)	(.56)	(.59)	(.75)
All internal causes	.39	1.07	1.69	1.25
	(.54)	(.80)	(.74)	(1.01)
Alcohol-related causes	.44	.80	.74	1.03
	(.21)	(.32)	(.33)	(.41)
Controls	age	age, age^2, interacted with over-21	age	age, age^2, interacted with over-21
Sample size	48	48	24	24

Notes: This table reports coefficients on an over-21 dummy from regressions of month-of-age-specific death rates by cause on an over-21 dummy and linear or interacted quadratic age controls. Standard errors are reported in parentheses.

Now All in One Table

TABLE 4.1
Sharp RD estimates of MLDA effects on mortality

Dependent variable	Ages 19–22 (1)	Ages 19–22 (2)	Ages 20–21 (3)	Ages 20–21 (4)
All deaths	7.66	9.55	9.75	9.61
	(1.51)	(1.83)	(2.06)	(2.29)
Motor vehicle accidents	4.53	4.66	4.76	5.89
	(.72)	(1.09)	(1.08)	(1.33)
Suicide	1.79	1.81	1.72	1.30
	(.50)	(.78)	(.73)	(1.14)
Homicide	.10	.20	.16	−.45
	(.45)	(.50)	(.59)	(.93)
Other external causes	.84	1.80	1.41	1.63
	(.42)	(.56)	(.59)	(.75)
All internal causes	.39	1.07	1.69	1.25
	(.54)	(.80)	(.74)	(1.01)
Alcohol-related	.44	.80	.74	1.03

Now All in One Table

Dependent variable	Ages 19–22 (1)	Ages 19–22 (2)	Ages 20–21 (3)	Ages 20–21 (4)
All deaths	7.66	9.55	9.75	9.61
	(1.51)	(1.83)	(2.06)	(2.29)
Motor vehicle accidents	4.53	4.66	4.76	5.89
	(.72)	(1.09)	(1.08)	(1.33)
Suicide	1.79	1.81	1.72	1.30
	(.50)	(.78)	(.73)	(1.14)
Homicide	.10	.20	.16	−.45
	(.45)	(.50)	(.59)	(.93)
Other external causes	.84	1.80	1.41	1.63
	(.42)	(.56)	(.59)	(.75)
All internal causes	.39	1.07	1.69	1.25
	(.54)	(.80)	(.74)	(1.01)
Alcohol-related causes	.44	.80	.74	1.03
	(.21)	(.32)	(.33)	(.41)
Controls	age	age, age^2,	age	age, age^2,

Non-Paramteric RDD

- The second way in which can handle non-linearities is by removing parametrical assumptions (about the slopes and how they change).
- This involves either taking simple averages, or computing linear regressions but only arround on a narrow bandiwth around the cut-off.
- This approach does not have the problems trying to get the relationship between a and $(\overline{M}_a)$ right, but it discard a large amount of data (information).
- The main challenge is how to choose the bandwidth to balance the trade of between bias (incorrectly attributing discontinuities) and variance (due to smaller sample size). The choice of this bandwidth is a judgement call, and results should not rely on one specific choice.
- It also has several "fancy" (more complex) methodological challenges that we ignore for now.

RDD: Final Considerations

- Visual inspection of RDD estimates are important but remember to keep an eye on the range of the y-axis
- Notice here that we cannot interpret the result of regression as a matched group, because we do not have individuals in the same cell (say age 20) with both treatment and control. The validity of RDD depends on our willingness to extrapolate across the running variable, at least around a narrow neighborhood around the cut-off.
- This extrapolation limits the policy questions that can be answered with RDD evidence. RDD can answer questions about changes in the margin (from 21 to 22 or 19) but not complete rearrangements of a policy (prohibiting or eliminating restrictions completely).

Acknowledgments

- MM