Regression Discontinuity Design

Introduction and Practical Advice

Liam Rose

Health Economics Resource Center, VA Palo Alto

Introduction

Regression discontinuity designs (RDDs or RDs) are a quasi-experimental design. With the right setup, the estimates are *causal*.

Today

- Fundamentals
- How to interpret
- How to implement

POLL QUESTION

What is your interest in regression discontinuity design?

- Reading and interpreting other papers
- Use on a project
- Expand analytic toolkit
- Just curious

Basics

Basics

- When RCTs are not feasible, often have many confounders, some observed and some not
    - This generates omitted variable bias
    - Adjustment on observables, matching, and machine learning cannot get around this issue

Basics

- Instead, use a threshold or cutoff to determine treatment status
  - Treatment = exposure to a policy
  - No treatment = no exposure to a policy
- Under the right circumstances, individuals will be very similar close to the threshold. However some of the individuals will not be treated and some will be treated.
- Comparing these very similar individuals around the threshold $\rightarrow$ We can get the causal effect of a policy/rule.

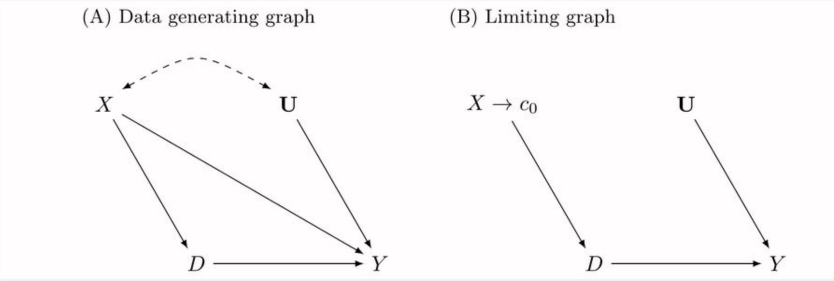


Figure 1: X (the thing we care about) is related to both the treatment (D) and the outcome (Y). With an RD, as X gets close to the cutoff, treatment and control units overlap, and X only affects Y through the treatment.

Example

- Policy = Legal access to alcohol.

Example

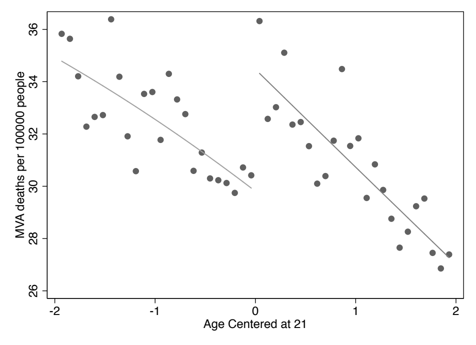
- Policy = Legal access to alcohol.
- Individuals just below 21 years are very, very similar to those just above 21 years. The only thing that differs between these groups is legal access to alcohol.

Example

- Policy = Legal access to alcohol.
- Individuals just below 21 years are very, very similar to those just above 21 years. The only thing that differs between these groups is legal access to alcohol.
- By comparing the mortality rate of individuals just below age 21 to those just above 21, with the RD approach, we can estimate the causal effect of alcohol access on mortality.
- Best shown with figures...

Treatment effect equals the jump or "discontinuity" in the graph at the threshold (21 years).

Figure 2: Mortality rate due to motor vehicle accidents



Basics

RD requires a lot of data and a specific cutoff, but luckily we tend to like rules and have many policies we can study.

- Can be applied to a lot of scenarios: School entry age, elections, test scores, newborn birth weight, and *Medicare eligibility age (65)*.

RD requires a lot of data and a specific cutoff, but luckily we tend to like rules and have many policies we can study.

- Can be applied to a lot of scenarios: School entry age, elections, test scores, newborn birth weight, and *Medicare eligibility age (65)*.
- It is attractive because of its simplicity: it is just OLS, and the figures tell the story

Basics

RD requires a lot of data and a specific cutoff, but luckily we tend to like rules and have many policies we can study.

- Can be applied to a lot of scenarios: School entry age, elections, test scores, newborn birth weight, and *Medicare eligibility age (65)*.
- It is attractive because of its simplicity: it is just OLS, and the figures tell the story
- Not event study/interrupted time series

Basics

Requirements:

- A continuous measure

Basics

Requirements:

- A continuous measure
- A cutoff that is:

Basics

Requirements:

- A continuous measure
- A cutoff that is:
    - non-manipulable

Basics

Requirements:

- A continuous measure
- A cutoff that is:
  - non-manipulable
  - arbitrary

Basics

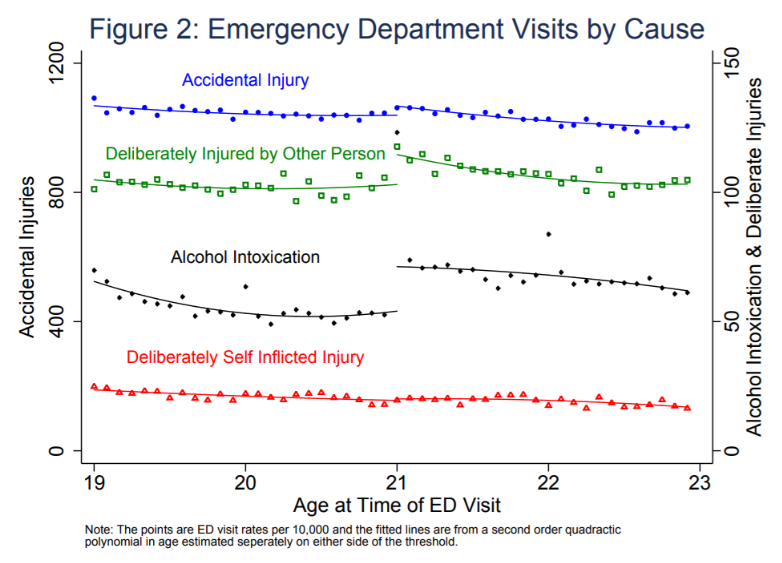
Requirements:

- A continuous measure
- A cutoff that is:
  - non-manipulable
  - arbitrary
- Data to test that these requirements hold

Interpretation

Let's look at some examples to be able to read and interpret RD estimates, starting with the Age 21 threshold.

Example - Age 21



Example - Age 21

	All Visits	Illness	Injury or Alcohol	Alcohol	Accidental Injury	Self Inflicted Injury	Injury Inflicted by Other
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
All							
Over 21	71.3	13.5	57.8	17.2	28.4	0.6	11.6
	(17.6)	(14.9)	(8.9)	(2.3)	(8.3)	(1.2)	(2.8)
Constant	3,973.8	2,758.2	1,215.6	54.1	1,039.0	19.5	103.1
	(16.1)	(13.8)	(6.6)	(1.3)	(7.1)	(1.1)	(2.1)
Observations	48	48	48	48	48	48	48
R-squared	0.927	0.961	0.845	0.914	0.781	0.602	0.697

- Estimates are on 'Over 21': This is the size of the jump
  - Column (4) - At age 21, there is an increase of 17.2 ED Visits per 10,000 people for alcohol intoxication
  - See figure, black line

Example - Age 21

	All Visits	Illness	Injury or Alcohol	Alcohol	Accidental Injury	Self Inflicted Injury	Injury Inflicted by Other
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
All							
Over 21	71.3	13.5	57.8	17.2	28.4	0.6	11.6
	(17.6)	(14.9)	(8.9)	(2.3)	(8.3)	(1.2)	(2.8)
Constant	3,973.8	2,758.2	1,215.6	54.1	1,039.0	19.5	103.1
	(16.1)	(13.8)	(6.6)	(1.3)	(7.1)	(1.1)	(2.1)
Observations	48	48	48	48	48	48	48
R-squared	0.927	0.961	0.845	0.914	0.781	0.602	0.697

- The constants give the "just before" estimate. Useful for interpretation.
    - Column (4) - For those almost age 21, there the ED visit rate of 54.1 per 10,000 people for alcohol intoxication

Example - Age 21

	All Visits	Illness	Injury or Alcohol	Alcohol	Accidental Injury	Self Inflicted Injury	Injury Inflicted by Other
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
All							
Over 21	71.3	13.5	57.8	17.2	28.4	0.6	11.6
	(17.6)	(14.9)	(8.9)	(2.3)	(8.3)	(1.2)	(2.8)
Constant	3,973.8	2,758.2	1,215.6	54.1	1,039.0	19.5	103.1
	(16.1)	(13.8)	(6.6)	(1.3)	(7.1)	(1.1)	(2.1)
Observations	48	48	48	48	48	48	48
R-squared	0.927	0.961	0.845	0.914	0.781	0.602	0.697

- The constants give the "just before" estimate. Useful for interpretation.
  - Column (4) - For those almost age 21, there the ED visit rate of 54.1 per 10,000 people for alcohol intoxication
  - This means that there is a ≈32 percent increase in ED visits for alcohol intoxication at age 21

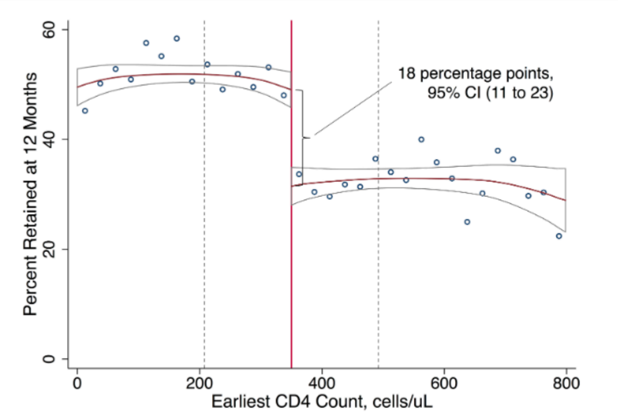


Figure 3: The proportion continuing clinical care after 12 months

Table 2. Intention-to-treat effects of ART eligibility on ART initiation and retention in HIV care.

Outcome	ART initiation by 6 months	Retained at 12 months (labs, ART, clinic visits)	Retained 0–6 months (labs, ART)	Retained 6–12 months (labs, ART)	Retained 12–18 months (labs, ART)	Retained 18–24 months (labs, ART)	Retained at 12 months (labs, ART)
Risk difference at 350-cells/μl CD4 threshold							
Regression coefficient	25.4	17.9	17.1	8.2	4.6	9.1	11.2
95% CI	(19.7, 31.1)	(11.4, 24.3)	(11.3, 22.9)	(3.8, 12.6)	(−1.0, 10.1)	(2.4, 15.8)	(4.2, 18.1)
p-Value	<0.001	<0.001	<0.001	<0.001	0.108	0.007	0.002
Predicted outcomes at 350-cells/μl CD4 threshold							
Eligible for ART (CD4 just below 350)	43.2	49.7	47.4	28.8	21.7	19.0	41.0
Not eligible for ART (CD4 just above 350)	17.8	31.8	30.3	20.6	17.2	9.9	29.9
IK bandwidth, cells/μl	96.4	142.1	114.2	164.7	125.4	164.2	116.8
N	3,354	3,327	3,937	5,478	2,954	1,734	2,733

Column (2) gives the estimate of the jump from the figure –> 17.9 percent more people at retained at 12 months, a 56 percent change

Implementation

What do we need?

Main requirements:

- A continuous measure (sometimes called the ”running” or ”forcing” variables)
- An arbitrary, non-manipulable cutoff
- A smooth distribution of characteristics besides the treatment at this threshold.

What do we do?

Plot it!

- You will need to make a few decisions here

Plot it!

- You will need to make a few decisions here
- Best practice is to plot everything. You want your results to relatively robust these choices

Plot it!

- You will need to make a few decisions here
- Best practice is to plot everything. You want your results to relatively robust these choices
- Start by plotting the running variable against the outcome

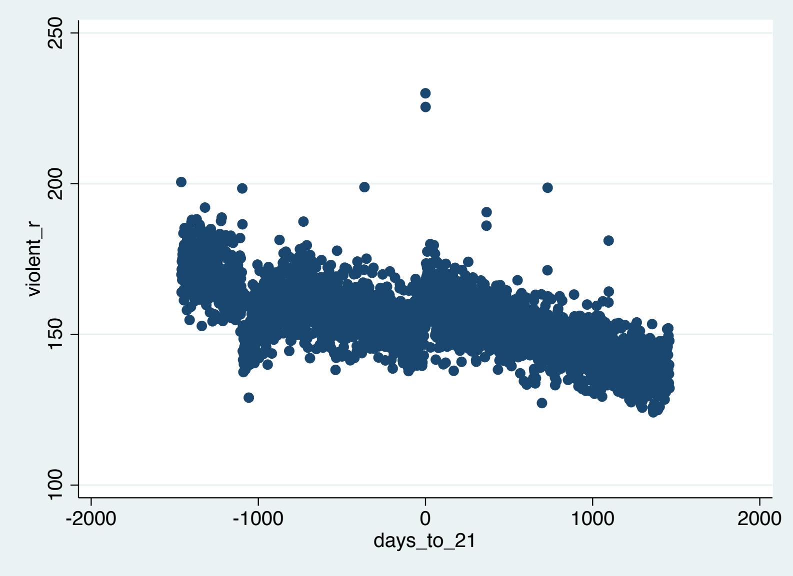


Figure 4: Violent crime rate relative to Age 21

Bins

- ”Bin” the data to make the figure clean
- Important: Make sure the bin does not span the threshold
- Do not run regressions on the binned data

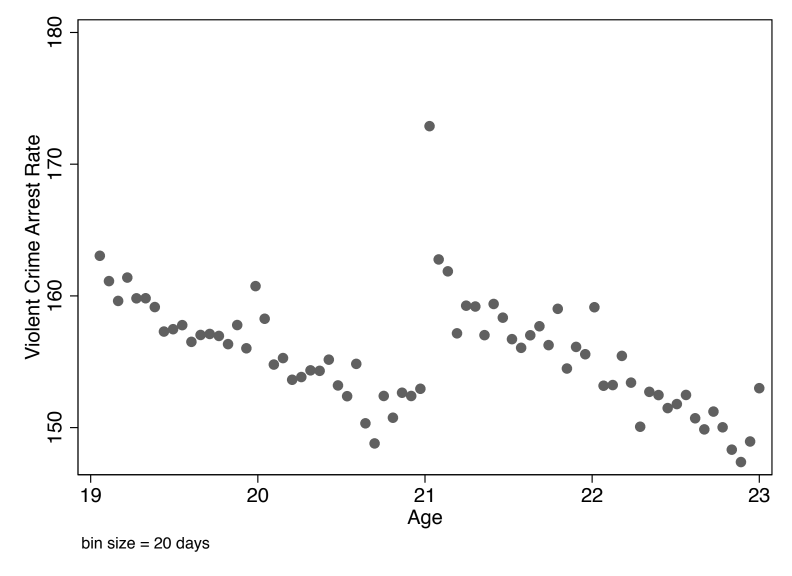
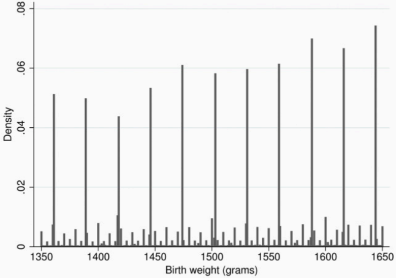


Figure 5: Violent crime rate relative to Age 21

Check to make sure the density is smooth

- This is called a McCrary Test [McCrary 2008]
- Want to make sure the distribution of the running variable is smooth across the threshold
    - Rounding/Measurement error can render otherwise valid applications invalid

Density Examples



(a) bad

(b) good

Bandwidth

Bandwidth is essentially just the range of x you include

- This is very important in this context

Bandwidth

Bandwidth is essentially just the range of x you include

- This is very important in this context
- Use optimal bandwidth procedures, but also try out a range

Bandwidth is essentially just the range of x you include

- This is very important in this context
- Use optimal bandwidth procedures, but also try out a range
    - Ideally, want a figure that shows robustness

Bandwidth

Bandwidth is essentially just the range of x you include

- This is very important in this context
- Use optimal bandwidth procedures, but also try out a range
    - Ideally, want a figure that shows robustness
    - Plot the RD estimate as a function of the bandwidth choice

Bandwidth Robustness

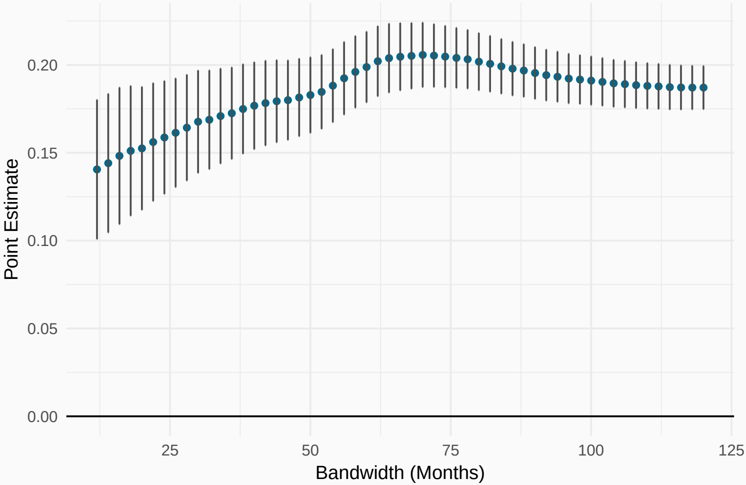


Figure 6: Estimates of the increase at the threshold. Each point is an estimate, and the lines are 95% confidence intervals

RD Balance

Provide evidence that the only thing changing at the threshold is the treatment

- This is similar to a balance check for an RCT
- To do this, simply change the outcome of interest to various characteristics
  - If these do not change at the threshold, we can be confident unobservables do not change either

RD Balance Example

Appendix Q: Change in Potential Confounders at Age 21

	Married	Employed	No HS Diploma	HS Graduate	Some College	Health Insurance
Estimated Change at Age 21	0.04 (2.54)	2.37 (5.34)	0.07 (3.53)	-0.09 (3.56)	1.06 (5.33)	2.24 (4.28)

Notes: See notes to Appendix N. The dependent variable for each regression is at the top of the column. The point estimates and their standard errors have been multiplied by 100 to convert to percents

Polynomial Order

We can use different polynomial orders on the regression lines, or use a local linear regression

- Once again, it is good to try out different choices
- However, too high is almost always a bad idea

Code Snippets and Packages

R

```
my_data <- read_csv("data")
my_date2 <- my_data %>%
   mutate(run = age - 21) %>%
   mutate(r2 = run^2) %>%
   mutate(r3 = run^3) %>%
   mutate(z = ifelse(run >=0,1,0) %>%
   mutate(interact1 = run*z) %>%
   mutate(interact2 = r2*z) %>%
   mutate(interact3 = r3*z)

lm(outcome ~ z + run + interact1 +
   r2 + interact2, data=mydata2)
```

Stata

```
use "data", clear
gen run = age - 21
gen run_sq = run^2
gen run_cu = run^3
gen z = 1 if run >= 0
replace post = 0 if run < 0
gen run_post = run*post
gen run_sq_post = run_sq*post
gen run_cu_post = run_cu*post

reg outcome z run run_post
    run_sq  run_sq_post
```

Packages in R: 'rdd', 'rdrobust'

Packages in Stata: 'rd', 'rdrobust'

Example - Time Permitting

- From Beland (2015) - Does the party of the state governor matter for black-white earnings gap?
- Data from elections and CPS
- Use close elections as RD

Beland (2015) - Governor Races

```r
library(tidyverse)
library(haven)
library(estimatr)
Political_laborSample1 <- read_dta("Documents/113580-V1/Politic
slimdata <- Political_laborSample1 %>%
  dplyr::select(black2, wages2, marginggg, totalhoursapp, wgt)
head(slimdata)
```

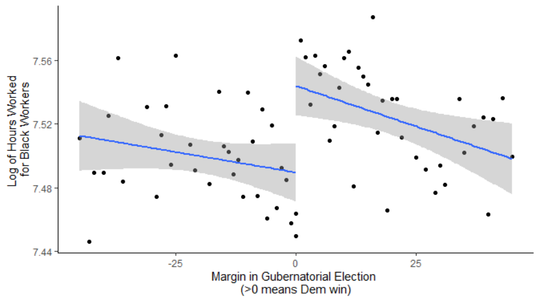
```
    > head(slimdata)
# A tibble: 6 x 5
  black2 wages2 marginggg totalhoursapp   wgt
   <dbl>  <dbl>     <dbl>         <dbl> <dbl>
1      0 11419.        45          1200 1632.
2      1      0        45             0 1718.
3      1      0        45             0 1493.
4      0  1649.        45           320 1329.
5      0      0        45             0 1426.
6      0      0        45             0 1437.
```

Beland (2015) - Governor Races

```
# reg1 <- lm_robust(totalhoursapp ~ z + marginggg
# + interact1, data = slimdata)
agg <- Political_laborSample1 %>%
  filter(black2 == 1) %>%
  filter(wages2 >0) %>%
  group_by(marginggg) %>%
  summarise(mean = weighted.mean(totalhoursapp ,wgt, na.rm=T))
```

Beland (2015) - Governor Races

```
agg %>%
  ggplot(., aes(x = marginggg, y= log(mean))) + geom_point() +
  stat_smooth(data = . %>% filter(marginggg <0), method = 'lm'
  stat_smooth(data = . %>% filter(marginggg >=0), method = 'lm
  xlab("Margin␣in␣Gubernatorial␣Election␣\n(>0␣means␣Dem␣win)"
  ylab("Log␣of␣Hours␣Worked␣\nfor␣Black␣Workers") +
  theme_classic()
```



- Make a lot of figures
- Try out different choices
- Show robustness

Conclusion and Questions

Conclusion

- RD gives a way to get causal estimates when an RCT is not feasible
- Leverage a continuous measure with an arbitrary cutoff to determine treatment
- Need to show:
    - Balance across the threshold
    - A smooth density
    - Robustness to parameters

QUESTIONS

Or via email: liam.rose@va.gov

Resources

- *Mastering 'Metrics* - masteringmetrics.com
    - Book and associated resources
    - *Mostly Harmless Econometrics* - more advanced version
- *Causal Inference: The Mixtape* - https://mixtape.scunning.com/
    - A brilliant guide to practical causal methods. The online version is free!

- Lee, David S., and Thomas Lemieux. (2010). "Regression discontinuity designs in economics." Journal of economic literature.
- Jacob, R., Zhu, P., Somers, M. A., & Bloom, H. (2012). A Practical Guide to Regression Discontinuity.