RDD and diff-in-diff

Silje Synnøve Lyder Hermansen

08-12-2020

Regression discontinuity design (RDD)

Basic assumption

RDD assumes a running variable (x) with a cut point (c) beyond which treatment is assigned (D).

$$D_i = \begin{cases} 1 & if \quad x_i \quad \geqslant c \\ 0 & if \quad x_i \quad < c \end{cases} \tag{1}$$

Distinction

It has a flavor of logit or propensity scores, but there are some differences:

- ~~**logit**~~ : x (not y) is not latent and we know the cutpoint: Both are observed and included as a *predictors*.
- ~~**matching**~~ : we have no control/treatment group. However, we assume that units on either side of the treatment are increasingly similar as their x is similar.

$\Rightarrow$ *Supposes clear rules with little administrative discretion.*

Examples

Administrative data are perfect: You have some rule that kicks in at a specific threshold for otherwise almost identical observations.

- school test scores on school admission, restrictions on class size
- legal drinking age on alcohol related deaths
- election of candidates in close races

Two ways of understanding RDD

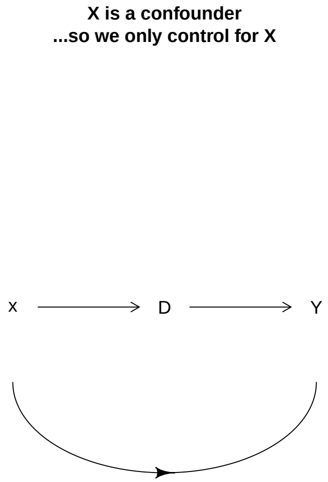
- **Individuals close to the threshold are interchangeable**

$\rightarrow$ in a small window, you have a treatment and a control group.

- x **is a bottleneck:** the relationship between D and Y is confounded by x, but all other confounders only influence Y through x.

$\rightarrow$ conditioning on x is sufficient to isolate the causal effect.

Two ways of understanding RDD

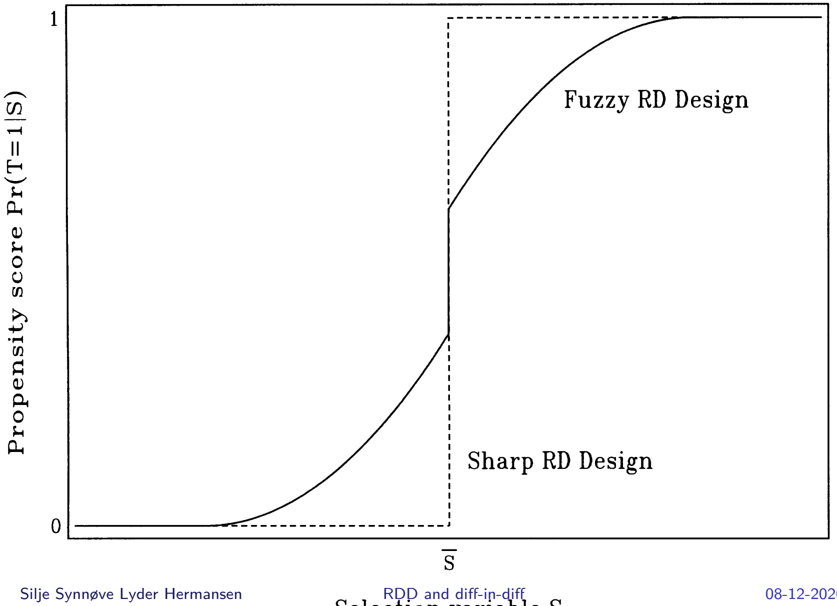


Two designs

We distinguish between two designs depending on how probable the treatment is:

- **sharp** RD: assignment is *deterministic*
- **fuzzy** RD: assignment is *probabilistic*

A visual representation



Sharp RDD

The basic model

The basic model

We assume the relationship between x and y is linear and the treatment is deterministic

$$y_i = \alpha + \rho \times D_i + \gamma \times x_1 + e_i \quad (2)$$

$\Rightarrow$ *The treatment is reported by* ρ

What is the treatment effect of legal drinking age (D) on deaths (y)?

```
##Load the data from my website: file df_ch4.rda
download.file(
  "https://siljehermansen.github.io/teaching/stv4020b/df_ch4.r
              "df_ch4.rda")
```

```
load("df_ch4.rda")

#Outcome: y
df$all

#Running variable x
df$age <- df$agecell - 21 #Center at cut point

#Recoding into a treatment variable D
df$over21 <- ifelse(df$agecell >= 21 , 1, 0)
```

What is the treatment effect of legal drinking age (D) on deaths (y)?

```
#Estimate the model
mod1 <- lm(all ~
             over21 + #Treatment (D)
             age,
           df)
```

What is the treatment effect of legal drinking age (D) on deaths (y)?

```
#Results
summary(mod1)
```

```
##
## Call:
## lm(formula = all ~ over21 + age, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0559 -1.8483  0.1149  1.4909  5.8043
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  91.8414     0.8050 114.083  < 2e-16 ***
## over21        7.6627     1.4403   5.320 3.15e-06 ***
## age          -0.9747     0.6325  -1.541     0.13
## ---
```

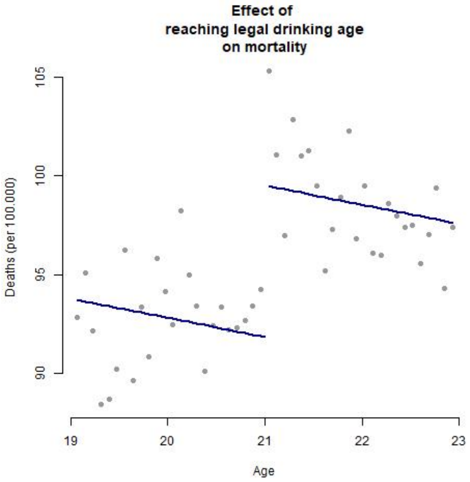
What is the treatment effect of legal drinking age (D) on deaths (y)?

```
#What is the treatment effect?
rho = mod1$coefficients["over21"]

paste("The treatment effect is",
      round(rho, 1),
      "more young people dying (per 100.000)",
      "when legal drinking age of 21 years is reached.")
```

```
## [1] "The treatment effect is 7.7 more young people dying (p
```

What is the treatment effect of legal drinking age (D) on deaths (y)?



Is that all?

This is true on two conditions

1. **no omitted variable bias**: x must capture all influence on D.
2. **the continuity assumption**: x must have a continuous effect on y

The continuity assumption

The continuity assumption

Sometimes we may pick up a smooth non-linear change by dummy coding

. . . that's not a regression discontinuity.

Ensuring linear effect

We can obtain a linear effect in two ways:

- recode the x $\rightarrow$ *parametric approach*
- consider a sufficiently small window $\rightarrow$ *non-parametric* approach

Ensuring linear effect: parametric approach

Recode the x

We can create a curvilinear effect of x using polynomials (e.g.:)

$$y_i = \alpha + \rho D_i + \gamma_1 x_i + \gamma_2 x_i^2 \tag{3}$$

```
df$age2 <- df$age^2
mod2 <- lm(all ~ over21 + age + age2,
           df)
```

Recode the x: polynomials

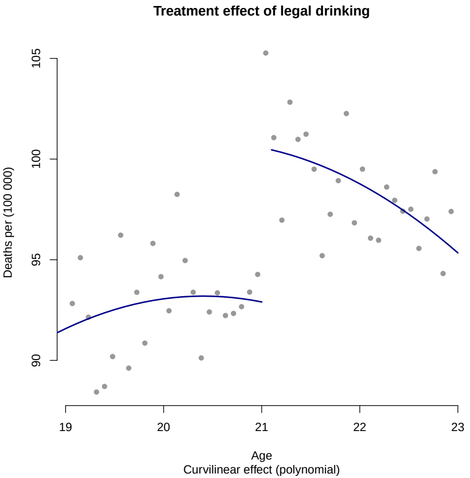
We can create a curvilinear effect of x using polynomials (e.g.:)

$$y_i = \alpha + \rho D_i + \gamma_1 x_i + \gamma_2 x_i^2 \tag{4}$$

```
df$age2 <- df$age^2
mod2 <- lm(all ~ over21 + age + age2,
           df)
```

$\Rightarrow$ *Here, x has a curved effect on both sides of the treatment.*

Recode the x: polynomial



Recode the x: interaction

We can assume x has different effects on each side of the treatment

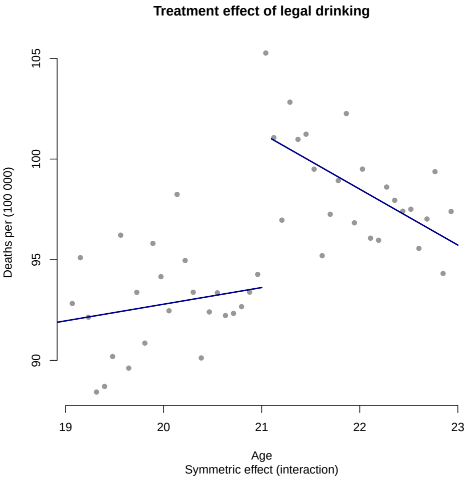
$$y_i = \alpha + \rho D_i + \gamma x_i + \delta x_i D_i \tag{5}$$

$\Rightarrow$ *we center the x on the cutpoint* $(x_i - c) \rightarrow \rho$ *still reports the change at the cutpoint.*

In R:

```
mod3 <- lm(all ~ over21 * age,
           df)
```

Recode the x: interaction



Extrapolation

We do this to estimate the effect at the cutpoint (ρ)

- but we can also extrapolate y beyond the cutpoint with x: $\rho + \delta(x - c)$

Ensuring linear effect: non-parametric approach

Bandwidth

- recoding the x is a *parametric approach*.
- subsetting the data to tweak the window around the cutpoint is a **non-parametric approach.**

$\Rightarrow$ *As the window becomes smaller, the shape of the* x*-effect matters less and less.*

Bandwidth : the idea

If the span of x around c is sufficiently small, there is no problem with non-linearity

- There's a tradeoff between linearity and statistical power (we need sufficient N).

Bandwidth : how do we choose it?

We try out different bandwidths

- We can do it by hand

```
mod4 <- lm(all ~ over21 + age,
           df[df$agecell > 20 & df$agecell < 22,])
```

$\Rightarrow$ *When you narrow down, do you get a weaker or stronger effect?*

Bandwidth : how do we choose it?

When narrowing down the window around the cutpoint, we'd expect a stronger effect.

Table 1:

	Dependent variable:	
	all	
	(1)	(2)
over21	7.663***	9.753***
	(1.440)	(1.934)
age	−0.975	−3.256*
	(0.632)	(1.700)
Constant	91.841***	91.713***
	(0.805)	(1.081)
Observations	48	24
R^2	0.595	0.703
Adjusted R^2	0.577	0.675
Residual Std. Error	2.493 (df = 45)	2.362 (df = 21)
F Statistic	32.995*** (df = 2; 45)	24.838*** (df = 2; 21)
Note:		*p<0.1; **p<0.05; ***p<0.01

⇒ *A trade-off with size of the data (and thus statistical power).*

Bandwidth : how do we choose it?

We try out different bandwidths

- We can do it by hand
- ...or we can make an algorithm do it:
  - run a local weighted regression line
  - bandwidth is estimated accordingly

$\Rightarrow$ *the point is to show robustness, not p-hack!*

Omited variable bias

We want to make certain that

- D has an effect on y :

→ is there really a cutpoint? Try out placebos!

- treatment was indeed assigned at the cutpoint:

→ is there unnatural clustering around one side?

- treatment has impact on outcome but not other pre-treatment covariates

→ check for balance/is there a similar “bump” for covariates? (bad news)

Fuzzy RD

Fuzzy RD

Often the D increases the probability of a treatment, but we don't know!

$\Rightarrow$ *This is a Instrumental Variable approach (more on Thursday)*

The exam school example

The exam school example: plan A

What's the effect of being around other good students on my 7th grade test scores?

- $y = \alpha + \beta_1 \bar{x} + \beta_2 x + e$
- $\bar{x}$: classmates' test scores in 4th grade (pre-treatment)
- x: my test scores in 4th grade

The exam school example: plan A

What's the effect of being around other good students on my 7th grade test scores?

- ~~$y = \alpha + \beta_1 \bar{x} + \beta_2 x + e$~~
- $\bar{x}$: classmates' test scores in 4th grade (pre-treatment)
- x: my test scores in 4th grade

$\Rightarrow$ *This is not a random assignment!*

The exam school example: plan B

Let's use the re-shuffeling due to exam schools.

- $y = \alpha + \beta_1 D + \beta_2 R + e$
- D: my assignment to a school
- R: my admittance exam results

$\Rightarrow$ *This is a even less random assignment!*

The exam school example: plan B

Let's use the re-shuffeling due to exam schools.

- ~~$y = \alpha + \beta_1 D + \beta_2 R + e$~~
- D: my assignment to a school
- R: my admittance exam results

$\Rightarrow$ *This is a even less random assignment! (although it is the reduced-form of plan C)*

The exam school example: plan C

Yes, let's use the re-shuffling due to exam schools.

- $\bar{x} = \alpha + \beta_1 D + \beta_2 R + e$

$\rightarrow$ predicted treatment assignment ($\tilde{x}$) as a function of my admittance scores (R) and the resulting admittance (or not) (D).

- $y = \alpha + \gamma\tilde{x} + \beta R + e$

$\rightarrow$ insert the part of classmate abilities due to my school admittance and control away my admittance scores (R)

$\Rightarrow$ *The treatment effect of classmates is expressed by γ!*

In brief

x **has a unique effect on** D**.** I'm interested in the effect of $\bar{x}$ on y, but x is completely endogenous:

- $y = \alpha + \phi\bar{x} + \beta_2 x + e$

I use treatment as an instrument. We do this in two steps

- step 1: $\bar{x} = \alpha_1 + \phi D + \beta_1 x + e_1$
- step 2: $y = \alpha_2 + \gamma\tilde{x} + \beta_2 x + e_2$

$\Rightarrow \gamma$ is the causal effect of D in a fuzzy design.

Differences-in-differences

Definition: Comparing two differences

Definition: Differences-in-differences

Treatment and control groups may differ in many ways (they are not randomly assigned)

- Pre-treatment: They move in parallel
- Post-treatment: They diverge

$\Rightarrow$ *Treatment effect is that difference*

$\Rightarrow$ *Assumes they would have otherwise continued in parallel*

What differences?

Diff-in-diff is based on two comparisons

- the difference pre- and post treatement *within* each unit
- the difference *between* the treatment and control groups

$\Rightarrow$ *based on panel data (units are observed several times).*

Example: States' monetary policy and number of banks

Take the differences between number of banks in two districts in Missisippi

- **Pre-treatment:** District 6 had 135 banks, while district 8 had 165.
- **Treatment:** District 6 provided money to banks, while district 8 did not.
- **Post-treatment:** After a year district 6 had 121 banks, while district 8 had 132

$\Rightarrow$ *What was the treatment effect?*

Example: States' monetary policy and number of banks

The two districts started out differently

- within-unit difference: Number of banks before and after the crisis in each district.

District 6: $121 - 135 =$ -14; District 8: $132 - 165 =$ -33

- between-unit differences: take the difference between the two.

$-14 - (-33) = 19$

$\Rightarrow$ *Basically a 2-by-2 table*

How to do it?

Interaction effects

In a regression, these differences are represented by an interaction term between two dummies

$$y_i = \alpha + \beta_1 T_i + \beta_2 P + \beta_3 T_i P_i \tag{6}$$

- P represents post-treatment effect: differences *within* units
- T represents the treatment group: differences *between* units
- β_3 is the causal effect

Data

Data requirements

- Requires panel data $\rightarrow$ which means correcting the standard errors.
- Common panel types: state-year/administrative unit-time period; people over time ...

$\Rightarrow$ *we want to know the trend before and after the break*

Another example: drinking age and death

Another example: drinking age and death

Does the legal drinking age has an effect on death rates among the young?

- y is number of deaths per 100 000
- P is post-treatment dummy
- T is dummies for states
- *trend* is year dummies

```
##Load the data from my website: file df_ch5.rda
download.file(
  "https://siljehermansen.github.io/teaching/stv4020b/df_ch5.r
  "df_ch5.rda")
```

Another example: step 1 → calculate differences

The authors have two tricks:

- Hardcode the interaction effect (dummy before/after treatment)
- They remove the intercept to retain all dummies

```
#Load the data
load("df_ch5.rda")

##with intercept
mod <- lm(mrate ~ legal +
            state +
            year_fct,
          df)
##without intercept; with all dummies
mod <- lm(mrate ~ 0 +
            legal +
            state +
            year_fct,
          df)
```

Another example: step 2 → calculate errors

Calculate robust standard errors:

```
library(clubSandwich)
```

```
## Warning: package 'clubSandwich' was built under R version 4
```

```
## Registered S3 method overwritten by 'clubSandwich':
##   method    from
##   bread.mlm sandwich
```

```
vcov <- vcovCR(mod, cluster = df[["state"]],
               type = "CR2")
robust <- coef_test(mod, vcov = vcov)$SE
```

Another example: step 3 → interpretation

Display the results and interpret:

```
library(stargazer)
stargazer(mod, se = robust,
                    omit = "state|year",
                    type = "html",
                    out = "regtable.html")
```

Another example: step 3 → interpretation

Table 2: Death rates among young as a function of legal drinking age

	Dependent variable:
	mrate
Legal drinking age (causal effect)	10.804**
	(4.479)
Observations	714
R^2	0.986
Adjusted R^2	0.985
Residual Std. Error	17.339 (df = 649)
F Statistic	726.005*** (df = 65; 649)
Note:	*p<0.1; **p<0.05; ***p<0.01

⇒ *What did we find?*

The parallel trends assumption

Main assumption

Units can be different, but – absent treatment – they must follow the same trend (hence the panel data).

- The regression assumes a counterfactual $\rightarrow$ remember the extrapolation.

Main assumption: The way around

When we have several treated and control units they can follow .

- individual trend lines. . .
- . . . that are modeled as deviations from one unique trend

Main assumption: The way around

When we have several treated and control units they can follow .

- individual trend lines...
- ... that are modelled as deviations from one unique trend

⇒ *We do that with an interaction effect!*

```
mod <- lm(mrate ~ 0 +
            legal +
            state *
            year_fct,
          df)
```

Last fix

If our units are in fact several units (say, populations in states)

- we can use weights

$\Rightarrow$ *There's a trade-off: treatment is at the unit level, statistical power at the subunit level.*