Statistics and Data science

Theory: Regression Discontinuity Design (RDD)

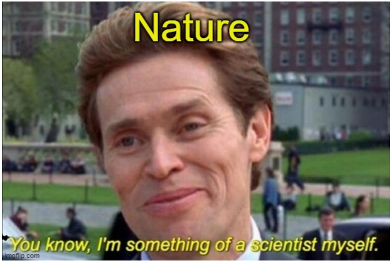
Q. Gallea[1]

[1]Enterprise 4 Society

Introduction

- RCT are not always possible.

- Hence, we will now focus on **natural experiments** :
  - *"A natural experiment is an observational study in which an event or a situation that allows for the random or seemingly random assignment of study subjects to different groups is exploited to answer a particular question."* Britannica
  - RDD, DiD, Synthetic controls, (IV)



Overview

1. [week 1] The gold-standard : RCT, A/B testing
2. **[today] Regression discontinuity design (RDD)**
3. [week 3] Difference-in-Difference (DiD)
4. [week 4] Synthetic controls

Overview

- In nature, boundaries or evolution through time tends to be smooth/continous. & space
  - Deserts and icecaps are far away
  - Climate change is a "slow" process
  - You don't become an adult overnight (or an expert in statistics)
  - ⇒ Difficult to compare those pairs of situations because they are far away (space/time) and hence other things vary in combination.

- **RDD key idea :** Find a discontinuity (often administrative rules). Then, treatement and control will be "close" and hence comparable (counterfactual).

Identifying assumption

The concept of discontinuity

- In nature, boundaries or evolution through time tends to be smooth/continous.
  - Desert and icecaps are far away
  - Climate change is a "slow" process
  - You don't become an adult overnight or a expert in statistics
  - ⇒ Difficult to compare the two because far away (space/time) and hence other things vary in combination.

- **RDD key idea :** Find a discontinuity. Then, treatement and control will be "close" and hence comparable (counterfactual).

Identifying assumption

Example

- Example : Does alcohol consumption increases the mortality rate ?

Identifying assumption

Example

- Example : Does alcohol consumption increases the mortality rate ?
  - **Naive comparison :** $E[DeathRate_i|Drink_i = 1] - E[DeathRate_i|Drink_i = 0]$
  - with i for age group.

Identifying assumption

Example

- Example : Does alcohol consumption increases the mortality rate ?
  - **Naive comparison :** $E[DeathRate_i|Drink_i = 1] - E[DeathRate_i|Drink_i = 0]$
  - with i for age group.

  $\Rightarrow$ Children usually are not consuming alcohol (low mortality on average in developed countries)

  $\Rightarrow$ Sick people might be advised to stop drinking alcohol (people with high death risk/comorbidity)

  $\Rightarrow$ Numerous confounding factors and hence $E[DeathRate_{0,i}|Drink_i = 0] \neq E[DeathRate_{0,i}|Drink_i = 1]$

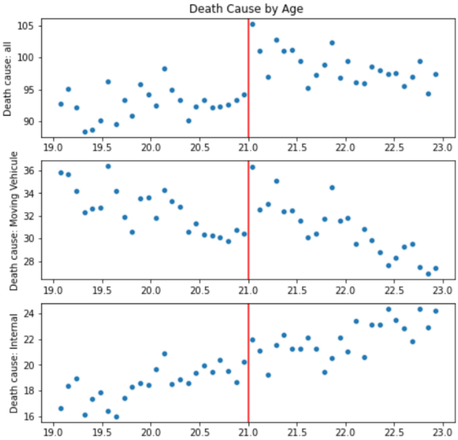
Identifying assumption

Example

- Example : Does alcohol consumption increases the mortality rate ?
  - **Naive comparison :** $E[DeathRate_i|Drink_i = 1] - E[DeathRate_i|Drink_i = 0]$
  - with i for age group.
  - ⇒ Children usually are not consuming alcohol (low mortality on average in developed countries)
  - ⇒ Sick people might be advised to stop drinking alcohol (people with high death risk/comorbidity)
  - ⇒ Numerous confounding factors and hence $E[DeathRate_{0,i}|Drink_i = 0] \neq E[DeathRate_{0,i}|Drink_i = 1]$
- **BUT :** Discontinuity : legal drinking age

Identifying assumption

Example



Identifying assumption

Rubin Causal Model

- Recall that :

$$E(Y_i|D_i=1)-E(Y_i|D_i=0)=$$
$$E(Y_{i1}|D_i=1)-E(Y_{i0}|D_i=1)+\underbrace{[E(Y_{i0}|D_i=1)-E(Y_{i0}|D_i=0)]}_{\neq 0}$$

- **Identifying assumption :**
$\lim_{\delta\to 0} E(Y_{0i}|X_0<X_i<X_0+\delta)-E(Y_{0i}|X_0-\delta<X_i<X_0)=0$
- with X_0 a threshold where the treatment changes (if $X_i>X_0 \Rightarrow D_i=1$ else $D_i=0$)
- The closer you get to the threshold (the discontinuity) the more the observations are similar (on observables and non-observables).

Always take data around the discontinuity -> not just at the one age

Identifying assumption

Rubin Causal Model

- Hence, under this assumption we are able to measure the causal effect.

$$\lim_{\delta \to 0} E(Y_i|X_0 < X_i < X_0 + \delta) - E(Y_i|X_0 - \delta < X_i < X_0) = E(Y_{1i} - Y_{0i}|X_i = X_0)$$

- $E(Y_{1i} - Y_{0i}|X_i = X_0)$
- Causal effect measured for the individuals close to the threshold.

Identifying assumption

Alcohol

- A 5-year-old (not drinking) is not comparable to a 76-year-old person who drinks alcohol.
- However, 20.5-year-old vs. 21.5-year-old people are highly similar, but the latter can legally buy alcohol in the US.

- **Assumptions :**
  - No other discontinuity at this age (any idea ?) maybe finishing college
  - Diseases risks, behavior, etc. move relatively smoothly making the two groups comparable.
  - In other words, the group of 20.5 year old is a good counterfactual.

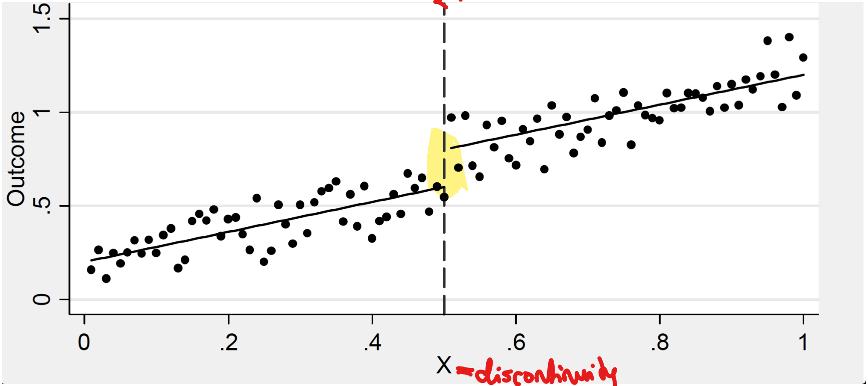
Linear regression for RDD

- $Y_i = f(X_i) + \beta D_i + \epsilon_i$
- with $f(X_i)$ any continuous fonction of X_i (linear in parameters)
- The idea is to capture the discontinuity, here β.
- It is very simple to implement. A simple linear regression. But it's also flexible depending on $f(.)$ and the inclusion of fixed effects and controls.

Functional form : linear

$$Y_i = \alpha X_i + \beta D_i + \epsilon_i$$

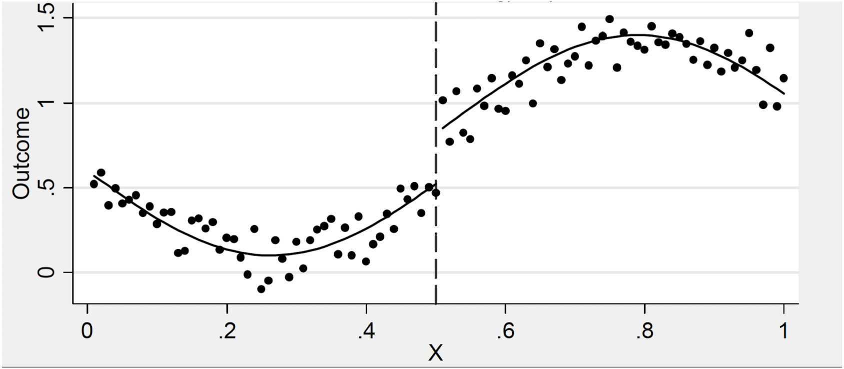
21



= discontinuity

Functional form : non-linear

$$Y_i = f(X_i) + \beta D_i + \epsilon_i$$



Functional form : Sharp vs. Fuzzy

Sharp RD

- **Example : Effect of scholarship on income later in life.**
- The probability to get the treatment goes from 0 to 1 at the threshold (e.g. Scholarship based on SAT scores).

$$D_i = 1 \text{ if } X_i \geq c, \text{ 0 otherwise}$$

- X could be correlated with the outcome (high SAT score, more competent and hence higher income later in life).
- This is why we should control for X (using the correct functional form)

Functional form : Sharp vs. Fuzzy

Fuzzy RD Regression discontinuity -> IV must be exogenous in order to understand the risk

- The probability to get the treatment does not go from 0 to 1 at the threshold but the probability of getting the treatment increase discontinuously.
- **Example :** If you get a perfect GRE score, your probability to be accepted for a PhD program is higher.
- To estimate the effect with a Fuzzy RDD, we would need to use an Instrumental Variable.
- Unfortunately, this goes beyond this class.

Robustness

- Robustness tests allow seeing if the results are robust to different adjustments. It also allows challenging the identification assumption.
- Here are four types of tests that you should do after an RDD.

1. Alternative functional form for $f(X)$.
2. Choice of bandwith (distance to threshold).
3. Another discontinuity for other covariates (observed) ?
4. Sorting possible ?

for example there are more people getting a license at 21 -> can check with data

Robustness

Alternative functional form for $f(X)$.

- Try another functional form (linear, squared, or polynomial form)

a. Evaluate the fit (R^2, p-values)
b. And check the robustness (if the coefficient on D_i changes)

Robustness

Choice of bandwith (distance to threshold)

- Replicate with different bandwith

a. Evaluate the fit (R^2, p-values)
b. And check the robustness (if the coefficient on D_i changes)

Robustness

Another discontinuity for other covariates (observed) ?

- Run the model on control variables
- $Z_i = f(X_i) + \beta D_i + \epsilon_i$
- with Z_i a control variable

⇒ β statistically significant ? It could be caused by the effect on Y_i though. Hence it's not a deal-breaker but help to reflect and at least find a solid rational for the effect.

Robustness

Sorting

- Is it possible to manipulate the threshold ?
  1. Administrative rule really exogenous ?
  2. Subject able to manipulate their results ? For example for the GRE you can retake until you pass a certain threshold.

  ⇒ To detect this, you could look for a discontinuity in density :
    - Plot an historgram with equal sized bins on the X_i variable and check if there is a jump in density around the threshold.

Additional readings :

- Causal Inference for the Brave and True :

  https://matheusfacure.github.io/python-causality-handbook/16-Regression-Discontinuity-Design.html

- The Causal Mixtape :

  https://mixtape.scunning.com/06-regression_discontinuity