

# Class 18 Natural Experiment I: Regression Discontinuity Design

Dr. Wei Miao

UCL School of Management

November 27, 2024

## Section 1

# Natural Experiment

# Class Objectives

- Concept of regression discontinuity design
- Estimation of causal effects using regression discontinuity design
- Application of regression discontinuity design in the business field

# From RCTs to Secondary Data

- RCTs are the gold standard of causal inference: In an RCT, the treatment is randomized and hence uncorrelated with any confounding factors, i.e.,  $cov(X, \epsilon) = 0$
- In practice, however, it can be challenging to implement a perfect RCT.
  - 1 Crossover and spillover effects;
  - 2 Costly in terms of time and money
- Therefore, we may want to exploit causal effects from existing secondary data. Besides the **instrumental variable** method, we can also investigate **natural experiments**.

# Comparison: RCT & Natural Experiment

## i Natural Experiment

A **natural experiment** is an event in which individuals are exposed to the experimental conditions that are determined by **nature** or **exogenous factors beyond researchers' control**. The process governing the exposures arguably **resembles randomized experiments**.

### RCT

- 1 Assignment of treatment is randomized by us
- 2 Treatment is under control by us
- 3 Primary data

### Natural Experiment

- 1 Assignment of treatment is randomized by nature
- 2 Treatment is not controlled by us
- 3 Secondary data

## Section 2

# Regression Discontinuity Design

# What is an RDD

- A **regression discontinuity design (RDD)** is a natural-experimental design that aims to determine the causal effects of interventions by identifying a **cutoff** around which an intervention is as if randomized across individuals.

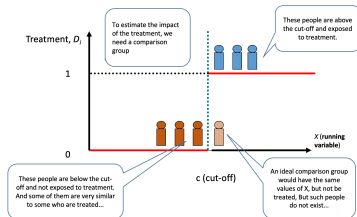


Figure 1: Visual illustration of RDD

## Motivating Example

**Business objective:** What is the causal effect of receiving a Master's degree with Distinction versus Non-Distinction on students' future salary?

- Can we run the following simple linear regression and obtain the causal effects?

$$Salary_i = \alpha + \beta Distinction_i + \epsilon_i$$

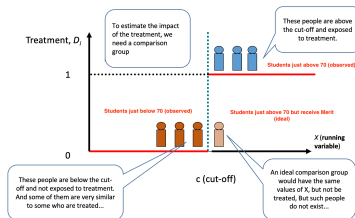
- Can we use RCT?
- Can we use instrumental variables?



# A Natural Experiment in the UK

- In the UK Education system, students receiving 70% or above final average grades will receive Distinction while students below 70% will receive Merit.
- The above setting gives us a nice natural experiment:
  - Students may improve their average grades significantly, such as moving from 60% to 69% by working harder, but they cannot perfectly control their average grades around the cutoff, say, from 69.9% to 70%.

# Visual Illustration of RDD: An Example of Distinction on Salary



## Why RDD Gives Causal Effects?

- For students just above 70%, to measure the treatment effects of receiving Distinction, we would need their *counterfactual salaries if they had not received Distinction*.
- At the same time, because the “running variable” **cannot be perfectly controlled** by the individuals **around the cutoff point**, it’s as if the treatment was randomized near the cutoff. Thus, individuals near the cutoff should be very similar, such that there should be no systematic differences across the treatment and control group.
  - Similar to RCT, we overcome the fundamental problem of causal inference using students just below 70 as the control group.
- All else being equal, a sudden change in the outcome variable at the cutoff can only be attributed to the treatment effect.

## Conditions for Using an RDD

- An RDD design arises when treatment is assigned based on whether an underlying **continuous variable** crosses a cutoff.
  - The continuous variable is often referred to as the **running variable**.
- **AND** the characteristic cannot be perfectly manipulated by individuals
  - We should only focus on individuals close to the cutoff point.

**Exercise:** eBay endorses sellers with 10,000 orders as Gold Seller. Can we use RDD to identify the causal effect of receiving Gold Seller endorsement on seller sales?

## Section 3

# Implementation of RDD

## Step 1: Select Sample of Analysis

- ① Determine the bandwidth above and below the cutoff and select the subset of individuals within the bandwidth
  - e.g., if we choose a bandwidth of 0.5, we need to filter out students with average scores between 69.5 and 70.5
- We face a trade-off when selecting the bandwidth: If we choose a smaller bandwidth around the cut-off
  - Pros: Individuals should be more similar around the cutoff, thus it is more likely the control group and treatment group are “as-if randomized”, thus higher internal validity.
  - Cons: We have a smaller subset of subjects which may not be representative of remaining individuals, thus lower external validity; We have a smaller sample size due to fewer individuals selected
- In practice, there is no specific rule how to determine the bandwidth. We need to run a set of different bandwidths as **robustness checks**.

## Step 2: Examine Continuity of Observed Characteristics

- ② Examine if other characteristics of the treatment group and control group are continuous at the cut-off point.
  - The idea is similar to “randomization check” in an RCT.

## Step 3: Data Analysis

- 3 Regress the outcome variable on the treatment indicator to obtain the causal effect.

$$Y_i = \beta_0 + \beta_1 Treated + \beta_2 running\_variable + \epsilon_i$$

- $Treated$  is a binary variable for whether or not the running variable is above the cutoff.
- We may also want to control the running variable in the regression to mitigate its confounding effects.



# The Causal Effect of Distinction on Salary

- We collect a dataset of 1000 graduates with their MSc final grade and salary.

	ID	score	experience	salary	Distinction
1	1	61.35711	0.3094854	61077.80	0
2	2	65.85353	0.4512458	62167.27	0
3	3	61.85750	0.8643365	60416.95	0
4	4	70.57253	2.5121579	68873.57	1
5	5	71.74155	0.6845075	68964.88	1

# Linear Regression Analyses

- Run a linear regression: `salary ~ Distinction`

```
feols(  
  fml = salary ~ Distinction,  
  data = data_rdd  
) %>%  
  modelsummary(  
    stars = T,  
    gof_map = c("nobs", "r.squared")  
  )
```

	(1)
(Intercept)	63306.835*** (57.722)
Distinction	5533.565*** (94.638)
Num.Obs.	1000
R2	0.774

+  $p < 0.1$ , \*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$

- The result suggests that Distinction can increase the salary by 5.5k, which is likely over-estimated due to omitted variable bias.

# RDD Analysis

- Step 1: Select a bandwidth around the cutoff, between 68% to 72%
- Step 2: Examine discontinuity of other variables (randomization check).
- Step 3: Run a linear regression on the subsample.

## RDD Results

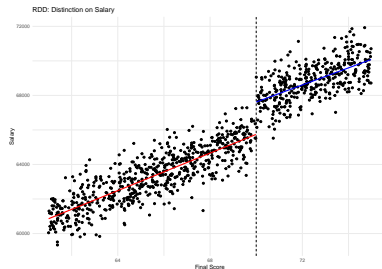
(1)	
(Intercept)	65201.101*** (81.291)
Distinction	2898.285*** (110.725)
Num.Obs.	282
R2	0.710

+  $p < 0.1$ , \*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$

- The result suggests that Distinction can increase the salary by 2.898k, which is likely a more accurate estimate of the causal effect than the OLS estimate.

# Visualization of RDD

```
pacman::p_load(ggplot2, ggthemes)
data_rdd %>%
  ggplot(aes(x = score, y = salary)) +
  geom_point() +
  geom_vline(xintercept = 70, linetype = "dashed") +
  labs(
    title = "RDD: Distinction on Salary",
    x = "Final Score",
    y = "Salary"
  ) +
  geom_smooth(
    data = subset(data_rdd, score < 70),
    method = "lm", se = FALSE,
    color = "red"
  ) +
  geom_smooth(
    data = subset(data_rdd, score >= 70),
    method = "lm", se = FALSE,
    color = "blue"
  ) +
  theme_minimal()
```



# Application of RDD in Marketing and Business Context

- **Scores vs. stars:** A regression discontinuity study of online consumer reviews: 4.49 rating vs. 4.5 rating result in a significant difference in sales.
- **The causal effect of Uber's surge pricing on demand and driver labor supply:** On the Uber platform, surge multiplier is rounded to the nearest 0.1, which creates a natural experiment for RDD.

## After-class Reading

- (recommended) [Quasi-experiment](#) (Econometrics with R)