Class 18 Natural Experiment I: Regression Discontinuity Design

Dr. Wei Miao

UCL School of Management

November 25, 2026

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 practice, however, it can be challenging to implement a perfect RCT.
  - Crossover and spillover effects
  - Costly in terms of time and money
- Therefore, we may want to exploit causal effects from existing secondary data. The **instrumental variable** method is a common way to estimate causal effects from secondary data. However, it requires finding valid instrumental variables, which can be challenging in practice.
- The **natural experiment** method is another way to estimate causal effects from secondary data, which is our focus in this lecture.
  - Regression discontinuity design (RDD)
  - Difference-in-differences (DiD)

Comparison: RCT & Natural Experiment

Natural Experiment

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

RCT

1. Assignment of treatment is randomised by us
2. Treatment is under control by us
3. Primary data

Natural Experiment

1. Assignment of treatment is randomised 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 randomised across individuals.

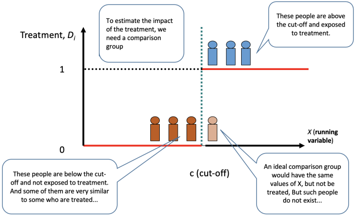


Figure 1: Visual illustration of RDD

Motivating Example

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

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

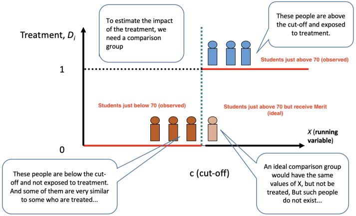
$$Outcome_i = \alpha + \beta Distinction_i + \epsilon_i$$

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

A Natural Experiment in the UK

- In the UK education system, students receiving a final average grade of 70% or above will receive a Distinction, while students below 70% will receive a 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 randomised 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 an 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 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 running variable cannot be perfectly manipulated by individuals

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

Data to Collect

- 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

- Key variables to collect:
  - Running variable: Final score
  - Outcome variable: Salary
  - Treatment variable: Distinction

Linear Regression Analyses

- Run a linear regression: `salary ~ Distinction`

	(1)
(Intercept)	63 306.835***
	(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 salary by 5533.565, which is likely over-estimated due to omitted variable bias.

RDD Analysis: Step 1

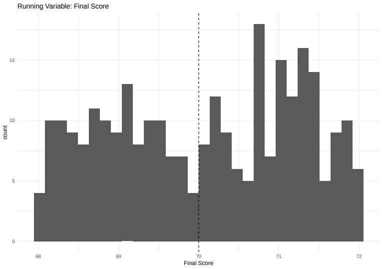
- Step 1: Select a bandwidth around the cutoff, between 68% to 72%

```
result_RDD <- feols(
    fml = salary ~ Distinction,
    data = data_rdd %>%
        filter(score > 68 & score < 72)
)
```

RDD Analysis: Step 2

- Step 2: Examine discontinuity of other variables (randomisation check).

```
# Visualise the running variable
data_rdd %>%
    filter(score > 68 & score < 72) %>%
    ggplot(aes(x = score)) +
    geom_histogram() +
    geom_vline(xintercept = 70, linetype = "dashed") +
    labs(
        title = "Running Variable: Final Score",
        x = "Final Score"
    ) +
    theme_minimal()
```



RDD Analysis: Step 3

- Step 3: Run the RDD regression within the bandwidth.

```
result_RDD <- feols(
    fml = salary ~ Distinction + score,
    data = data_rdd %>%
        filter(score > 68 & score < 72)
)
```

RDD Results

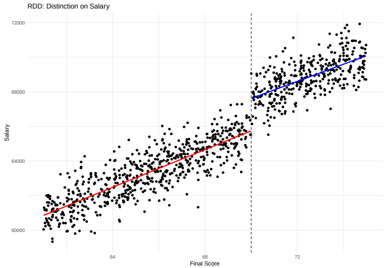
	(1)
(Intercept)	28 203.125***
	(6687.975)
Distinction	1777.295***
	(228.350)
score	536.677***
	(97.007)
Num.Obs.	282
R2	0.739

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

- The result suggests that Distinction can increase salary by 1777.295, which is likely a more accurate estimate of the causal effect than the OLS estimate.

Visualisation 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()
```



Additional Example

- Scores vs. stars: A regression discontinuity study of online consumer reviews:
  - eBay's star rating system creates a natural discontinuity
  - Sellers with average ratings 4.75 receive 5 stars; those below receive 4.5 stars
- Research question: Does the visual star rating affect sales?
  - **Treatment**: Receiving 5 stars vs. 4.5 stars at the 4.75 threshold
  - **Outcome**: Product sales and buyer behaviour
  - **Finding**: Significant jump in sales at the 4.75 cutoff, demonstrating causal effect of star visualisation

After-class Reading

- (recommended) Quasi-experiment (Econometrics with R)