Introduction to Causal Inference in R (Inchan Hwang)

Inchan Hwang, `inchan@yonsei.ac.kr`

4 October 2024

Introduction

Causal inference in social sciences primarily focuses on understanding relationships between variables. However, studying causation goes beyond mere correlations. It requires addressing confounding variables that might affect the observed relationship. There are two primary approaches to tackle this:

1. Randomization in treatment assignment.
2. Conditioning on confounders.

At the heart of causal inference lies the "What if" question:

- What if Thanos had not snapped his fingers?
- What if I majored in astronomy instead of sociology?
- What if I were born in Denmark instead of South Korea?

The core challenge is determining how outcomes would differ under alternate scenarios. This is the foundation of causal inference.

A concrete example of causal inference can be drawn from Thanos' actions. The treatment group comprises a world where half the population vanishes due to his finger snap, while the control group represents a world without the snap. Yet, in *Infinity War*, we cannot observe both potential outcomes simultaneously. This impossibility of observing both potential outcomes for the same individual unit is the **fundamental problem of causal inference**.

The solution? Design treatment groups that closely resemble the counterfactuals of the control group to minimize bias. This is the essence of identifying causal effects.

In summary, the need for assumptions in causal inference arises from the challenges of identifying causal effects and the data-generating process. Because we cannot observe different potential outcomes for the same individual, we seek to construct control groups that approximate the counterfactual world as closely as possible.

Next, we will discuss various levels of causal inference approaches:

- Randomized Controlled Trials (RCT)
- Regression Discontinuity
- Propensity Score Matching

---

1. Randomized Controlled Trials (RCT)

Overview

Randomized Controlled Trials (RCTs) are considered the gold standard for causal inference. By randomizing treatment assignment, we ensure that the treatment and control groups are comparable. This randomization minimizes selection bias and balances confounding variables across groups, providing a clean estimation of the treatment effect.

The key reason RCTs are effective for causal inference is that randomization ensures the independence of treatment assignment from potential outcomes. This independence eliminates confounding, allowing us to attribute differences in outcomes directly to the treatment. With RCTs, we can rely on simple statistical methods, such as t-tests, to estimate causal effects because randomization creates balance across groups.

Why a t-test Works in RCTs

In an RCT, the treatment and control groups are randomly assigned, ensuring that any differences in outcomes between the two groups are due to the treatment and not pre-existing differences. The t-test measures whether the mean outcome in the treatment group is statistically significantly different from the control group. Since randomization eliminates confounding, this difference directly represents the causal effect.

Why a t-test without Randomization May Not Work as a causal effects?

Without randomization, treatment assignment might be correlated with confounders—variables that influence both treatment and outcomes. For example, individuals with higher socioeconomic status might self-select into receiving a bachelor's degree, leading to a biased estimate of the treatment effect. In such cases, a t-test might capture both the treatment effect and the influence of confounders, making it invalid for causal inference.

Example: Simulating RCT Data

Before running the simulation, let's discuss the logic behind it:

- **Random Assignment**: Treatment is assigned randomly via a coin toss, ensuring there's no systematic difference between groups other than the treatment.
- **Outcome Generation**: Outcomes are generated with a baseline mean and some added noise. For the treatment group, a fixed effect is added to simulate the causal effect.
- **Analysis**: A t-test is used to compare the means of the treatment and control groups, and due to randomization, the observed difference can be interpreted as causal.

```r
# Simulate RCT data with coin toss
set.seed(123)

# Assign treatment based on coin toss
coin_toss <- sample(c("Heads", "Tails"), 100, replace = TRUE)
data <- data.frame(
  treatment = ifelse(coin_toss == "Heads", 1, 0),
  outcome = rnorm(n, mean = 10, sd = 2)
)

data$outcome[data$treatment == 1] <- data$outcome[data$treatment == 1] + 3
```

A random coin toss (`sample(c("Heads", "Tails"), n, replace = TRUE)`) assigns treatment (treatment = 1) or control (treatment = 0). An outcome variable is simulated (`rnorm(n, mean = 10, sd = 2)`) with a baseline mean of 10. The treatment group's outcome is increased by 3 (`data$outcome[data$treatment == 1] <- ... + 3`) to introduce a treatment effect.

```r
# Analyze the effect of treatment using t-test
t_test <- t.test(outcome ~ treatment, data = data)
print(t_test)
```

Explanation of Results

The t-test compares the mean outcomes of the treatment and control groups. Because treatment was randomly assigned:

- Any observed difference in means can be attributed to the treatment itself.
- Randomization ensures that confounders are equally distributed across groups, making this inference valid.

The t-test result provides both the magnitude of the difference (mean difference) and the statistical significance, helping us infer whether the treatment effect is real and not due to random chance.

Real-World Context and Results Interpretation

The Algebra I Initiative is a real-world educational reform designed to accelerate math learning for 9th-grade students classified as "below grade level." This example is based on the study: **"Accelerating Opportunity: The Effects of Instructionally Supported Detracking"** by Dee and Huffaker(2024).

Contextual Background:

- **Challenges Addressed:**
  - Math course stratification has historically been linked to inequities in ethnoracial and socioeconomic distributions.
  - "Tracking" students into remedial or advanced courses often reinforces segregation and diminishes opportunities for lower-performing students.
- **The Algebra I Initiative:**
  - Combined heterogeneous classroom settings with intensive teacher development.
  - Students typically assigned to remedial courses were placed directly into Algebra I, with additional teacher support to promote differentiated instruction.
  - Teachers participated in professional development, gaining strategies to support a mixed-ability classroom effectively.

Simulated Data and Relevance:

- The example simulation mirrors the real-world intervention by assigning treatment and control groups randomly.
- The outcome variable, standardized test scores (e.g., SBAC), reflects the focus on measurable academic improvements in the study.

Key Findings from the Study:

1. **Academic Achievement:**
   - Students "below grade level" saw an increase of +0.19 SD in 11th-grade math test scores, demonstrating substantial academic improvement.
   - This effect persisted over time, indicating the program's durability in improving outcomes.
2. **Course Progression:**
   - Treated students were more likely to complete advanced courses like Algebra II by the end of high school, reflecting enhanced academic trajectories.
3. **Student Engagement:**
   - A reduction in absenteeism and chronic absenteeism for "below grade level" students assigned to the initiative.
   - Improved district retention rates, highlighting a stronger sense of belonging and satisfaction within the educational environment.

Interpretation of Simulation Results:

- The simulated t-test reflects differences in post-treatment scores between treatment and control groups.
- In the real-world context:
  - **Treatment Effect Validity:** Randomization ensures that observed differences are attributable to the initiative.
  - **Practical Implications:** The standardized effect size of +0.19 SD is substantial, translating into meaningful improvements in long-term educational and economic outcomes for low-proficiency students.

By incorporating these contextual findings and interpreting the simulation results within this framework, the example illustrates the potential of RCTs to inform policy decisions that promote equity and achievement in education.

Example: Real-World Inspired RCT Data

This example simulates a randomized controlled trial (RCT) based on a study of Algebra I Initiative programs targeting 9th-grade students classified as "below grade level." Treatment and control groups are randomly assigned, and the impact of the program is assessed on outcomes such as standardized test scores.

Context:

- **Treatment Group**: Participated in the Algebra I Initiative, receiving differentiated support and professional development for teachers.
- **Control Group**: Followed the existing educational framework, including remedial pre-algebra courses.
- **Outcome Variable**: Standardized math test scores in 11th grade (e.g., SBAC scores).

Simulation:

The dataset simulates 200 students, randomly assigned to treatment and control groups:

```r
# Simulate Algebra I Initiative data
set.seed(456)
n <- 200

# Random assignment
students <- data.frame(
  treatment = sample(c(0, 1), n, replace = TRUE),
  pretest_score = rnorm(n, mean = 240, sd = 20)
)

# Generate post-treatment outcomes
students$posttest_score <- students$pretest_score +
  ifelse(students$treatment == 1, rnorm(n, mean = 15, sd = 10), rnorm(n, mean = 5, sd = 10))

# Analyze the effect of treatment using t-test
treatment_effect <- t.test(posttest_score ~ treatment, data = students)

# Display results
treatment_effect
```

Explanation of Results:

- **Treatment Effect**: The t-test compares post-test scores between the treatment and control groups.
- **Causal Inference**: Random assignment ensures that any observed differences can be attributed to the Algebra I Initiative. t_test <- t.test(outcome ~ treatment, data = data)

Explanation of Results

The t-test compares the mean outcomes of the treatment and control groups. Because treatment was randomly assigned:

- Any observed difference in means can be attributed to the treatment itself.
- Randomization ensures that confounders are equally distributed across groups, making this inference valid.

The t-test result provides both the magnitude of the difference (mean difference) and the statistical significance, helping us infer whether the treatment effect is real and not due to random chance.

2.2 Regression Discontinuity (RD)

Regression Discontinuity (RD) design is a quasi-experimental approach used to estimate causal effects when treatment assignment is determined by a cutoff or threshold on a continuous variable, known as the "running variable." This method is particularly useful in settings where randomized experiments are not feasible but a strict, rule-based mechanism determines who receives the treatment.

A quasi-experiment, unlike a true experiment, does not rely on full randomization. Instead, it takes advantage of naturally occurring or policy-driven conditions—such as a funding threshold, policy age limit, or score cutoffs—to approximate experimental conditions. RD is especially powerful for estimating causal effects in these settings because it focuses on units near the cutoff, where treatment assignment approximates randomization.

In RD, eligibility for treatment is determined by whether the running variable exceeds or falls below a predetermined cutoff. For instance, students scoring above 100 on a standardized test might qualify for additional tutoring, while those scoring below do not. Units just above and below the cutoff are assumed to be comparable, allowing for valid causal inference.

The key features of RD include:

1. **Running Variable:** A continuous measure that determines treatment eligibility, such as test scores or income levels.
2. **Cutoff or Threshold:** The value of the running variable that determines treatment status. For example, students scoring 100 may qualify for a program.
3. **Local Comparison:** Comparisons are made between units just above and below the cutoff, where differences in outcomes are attributed to the treatment.
4. **Sharp vs. Fuzzy RD:**
    - In sharp RD, the treatment is strictly assigned based on the cutoff.
    - In fuzzy RD, treatment assignment is probabilistic near the cutoff.

RD is commonly used in policy evaluations, such as determining the impact of scholarships, grants, or other programs awarded based on thresholds. Its advantage lies in its ability to provide credible causal estimates when randomized controlled trials are not possible. However, the validity of RD relies on several assumptions:

- The running variable must not be manipulable near the cutoff to ensure comparability.
- Units near the cutoff must be similar in all respects except for treatment status.
- Proper functional forms (e.g., linear or quadratic) must be used to model the relationship between the running variable and the outcome.

Mathematically, RD can be represented as:

$$Y_i = \beta_0 + \beta_1 T_i + f(X_i) + \epsilon_i$$

where Y_i is the outcome, T_i indicates treatment (1 if above cutoff, 0 otherwise), X_i is the running variable, and $f(X_i)$ is a smooth function of X_i (e.g., linear or polynomial).

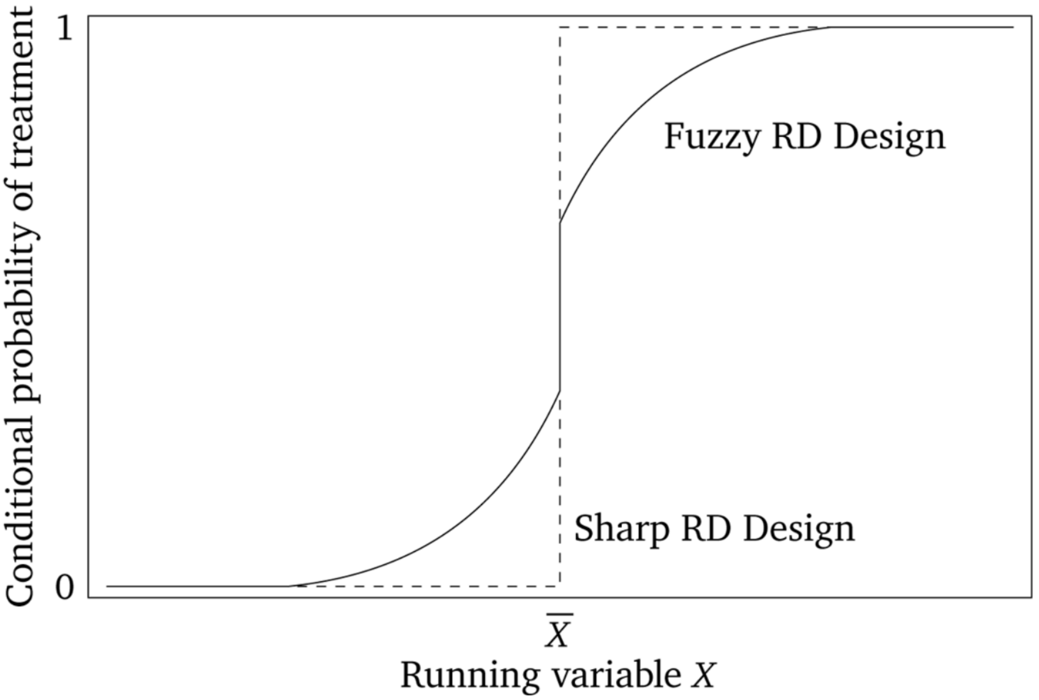

Figure 1: alt text

Real-World Example: California Career Pathways Trust (CCPT)

The California Career Pathways Trust (CCPT) program offers a compelling real-world example of RD in action. This program provided grants to school districts to develop career pathways aligned with local labor market needs. Treatment assignment was based on a strict application score cutoff, creating an ideal setting for RD analysis.

Context

- **Treatment:** School districts receiving CCPT grants used funding to develop career-oriented curricula, purchase equipment, and establish partnerships with community colleges and businesses.
- **Outcome of Interest:** The primary outcome studied was the school district dropout rate.
- **Assignment Mechanism:** Grant eligibility was determined by application scores, with a predetermined cutoff (e.g., scores 100).

Simulated Data and Analysis The following simulation mirrors the CCPT grant study, using application scores as the running variable and dropout rates as the outcome:

```r
# Simulate RD data for CCPT example
set.seed(123)
n <- 814

# Generate running variable (application scores)
data <- data.frame(
  running_var = runif(n, 50, 150),
  dropout_rate = rnorm(n, mean = 0.02, sd = 0.005)
)

# Treatment assignment based on cutoff
cutoff <- 100
data$treatment <- ifelse(data$running_var >= cutoff, 1, 0)

# Introduce treatment effect on dropout rates
data$dropout_rate[data$treatment == 1] <- data$dropout_rate[data$treatment == 1] - 0.005

# Fit RD model
library(rdd)
rd_model <- RDestimate(dropout_rate ~ running_var, data = data, cutpoint = cutoff)
summary(rd_model)
```

Interpretation of Results The output from the RDestimate function provides several estimates based on varying bandwidths around the cutoff. These include:

1. **LATE (Local Average Treatment Effect):** This represents the estimated causal effect of the treatment at the cutoff point. It focuses on units closest to the cutoff and uses the default bandwidth for the analysis. LATE is particularly useful for understanding the impact of the treatment where the running variable just meets or exceeds the threshold.

2. **Half-Bandwidth Estimate:** This estimate narrows the range around the cutoff by halving the default bandwidth. By focusing on fewer observations closer to the cutoff, it reduces potential bias from including observations farther from the threshold but increases variance due to fewer data points.

3. **Double-Bandwidth Estimate:** This estimate expands the range around the cutoff by doubling the default bandwidth. It includes more observations, which can reduce variance but may introduce bias by incorporating units farther from the cutoff that might not be directly comparable.

These estimates collectively help evaluate the robustness of the RD analysis. By comparing results across different bandwidths, researchers can assess how sensitive the estimated treatment effects are to the choice of bandwidth.

Key Takeaways from the Results

- **Bandwidth Sensitivity:** The results show how sensitive RD estimates can be to the choice of bandwidth. Narrower bandwidths reduce bias but increase variance, while broader bandwidths include more observations but risk introducing bias.
- **Significance Levels:** None of the estimates achieve statistical significance, meaning there is insufficient evidence to conclude a causal effect of treatment in this simulation. This could be due to noise in the data or the need for more refined functional forms.
- **Diagnostics and Warnings:** The warnings about singular covariances and NaNs in standard errors suggest potential issues with model specification or data distribution near the cutoff. It's crucial to perform additional robustness checks, such as varying the functional forms of the running variable or excluding outliers near the cutoff.

Practical Implications

While this simulation may not yield statistically significant results, it highlights the mechanics of RD analysis and the importance of rigorous diagnostic checks. In real-world scenarios, careful consideration of bandwidth, functional form, and data quality is essential to ensure credible causal estimates.

The RD analysis shows that districts receiving the CCPT grant (above the cutoff) experienced significantly lower dropout rates compared to those that did not receive the grant. This aligns with findings from the real-world study, which reported a 23% reduction in dropout rates for grant recipients. Key assumptions for the validity of these results include:

1. Districts just above and below the cutoff are comparable.
2. The scoring mechanism is not subject to manipulation.
3. Functional forms used in the analysis appropriately capture the relationship between the running variable and outcomes.

Practical Implications The CCPT grant illustrates how targeted funding can effectively reduce dropout rates and improve educational outcomes. RD analysis provides robust evidence of these effects, demonstrating the value of investing in career and technical education programs.

By implementing RD designs, researchers and policymakers can draw credible causal conclusions in settings where randomization is not feasible.