

경영경제데이터분석

RDD (Regression Discontinuity Design)

최 현 홍

(hongchoi@khu.ac.kr)

Contents

- Introduction
- What is RDD?
- Conducting RDD Analysis
- Comparing RDD with Other Methods
- Recap

Objectives

학습 목표

- RDD의 주요 개념을 이해한다
- RDD의 주요 장단점을 이해한다
- RDD의 설계 및 적용 방법을 이해한다
- RDD의 분석 결과를 적절히 해석할 수 있다

Introduction

Introduction

온라인 리뷰 평점

- 온라인 플랫폼을 통해 다양한 제품과 서비스를 구매할 수 있게 되면서, 소비자 의견을 표현하기 위한 온라인 리뷰를 제공하는 플랫폼 역시 많이 등장하고 있음

coupang

쿠팡에서 '커피 원두'를 검색한 결과

1



쿠팡추천
워너빈 시그니처 블렌디드 원두커피, 1kg, 1개, 홀빈(분쇄안함)
17,950원 로켓와우
(10g당 180원)
내일(목) 새벽 도착 보장
무료배송 · 무료반품
★★★★★ (9348)
최대 180원 적립

2



곰곰 콜롬비아 블렌드, 홀빈(분쇄안함), 1000g, 1개
23% 26,900
21,990원 로켓와우
(10g당 220원)
내일(목) 새벽 도착 보장
무료배송 · 무료반품
새 상품, 반품-미개봉 (2) 최저21,330원
★★★★★ (29915)
최대 220원 적립

3



스타벅스 하우스 블렌드 홀빈, 1.13kg, 1개
7% 42,570
39,390원 로켓와우
(10g당 349원)
내일(목) 새벽 도착 보장
무료배송 · 무료반품
★★★★★ (13591)
최대 394원 적립


 Tripadvisor

Tripadvisor에서 '홍콩 딤섬집'을 검색한 결과


Hong Kong Dim Sum

281 results match your filters [Clear all filters](#) Sort by: Relevance

Dim Sum X



1. Tin Lung Heen
MICHELIN
1,756 reviews · Closed Now
Chinese, Asian · \$\$\$
Elevated Cantonese cuisine with a focus on dim sum, traditional garapua, and Iberian pork. The setting provides panoramic views and a menu where the cha siu and BBQ pork buns are highlighted.
Reserve



2. Yung's Bistro
95 reviews
Chinese, Healthy · \$\$-\$\$\$
Cantonese cuisine with a creative twist, this restaurant including roasted goose, barbecued pork, and shrimp.
Order online

Introduction

리뷰 평점의 표현 방식



- 그런데, 많은 리뷰 플랫폼들의 경우 **5점 만점의 별점 체계**를 이용하여 **0.5점 단위로 제품에 대한 만족도를 표시**하도록 하는 경우가 많음
 - 그런데, 위 체계를 이용하여 평균 평점을 시각화함에 있어, **실제 평균 수치는 0.5 단위가 아니지만 시각화는 0.5 단위로 제시**하는 경우가 많음 (**반올림 컷오프**)

★★★★★ (9348) 커피 A

●●●●● 1,756 reviews **딴섬집 A**

★★★★★ (29915) 커피 B

●●●●● 95 reviews **딴섬집 B**

★★★★★ (13591) 커피 C

여러분이라면 어떤 커피를 구매하고 어느 딴섬집에 가겠는가?



Introduction

리뷰 평점의 효과

- 그런데, 많은 리뷰 플랫폼들의 경우 **5점 만점의 별점 체계**를 이용하여 **0.5점 단위로 제품에 대한 만족도를 표시**하도록 하는 경우가 많음
 - 그런데, 위 체계를 이용하여 평균 평점을 시각화함에 있어, **실제 평균 수치는 0.5 단위가 아니지만 시각화는 0.5 단위로 제시**하는 경우가 많음 (**반올림 컷오프**)

4.26 ★★★★★ (9348) 커피 A

4.34 ●●●●● 1,756 reviews 딴섬집 A

4.74 ★★★★★ (29915) 커피 B

4.99 ●●●●● 95 reviews 딴섬집 B

4.76 ★★★★★ (13591) 커피 C

실제 별점 수치를 보니 생각이 바뀌었는가?



Introduction

컷오프기반 별점 부여 방식에 따른 효과 분석

- Anderson & Magruder (2012)는 **RDD**를 활용하여 이러한 반올림 컷오프 기반 별점의 효과를 분석하였음
 - Anderson, M., & Magruder, J. (2012). Learning from the crowd: Regression discontinuity estimates of the effects of an online review database. The Economic Journal, 122(563), 957-989.
- 해당 연구는 미국에서 주로 활용되는 다양한 비즈니스 및 서비스에 대한 리뷰 공유 온라인 플랫폼인 **Yelp**의 식당 평점을 대상으로 하였음
 - 식당 뿐 아니라 미용실, 헬스 클럽, 병원 등에도 평점을 매길 수 있으며, 웹사이트 혹은 모바일 어플리케이션을 통해 온라인 예약, 배달 등도 가능



Introduction

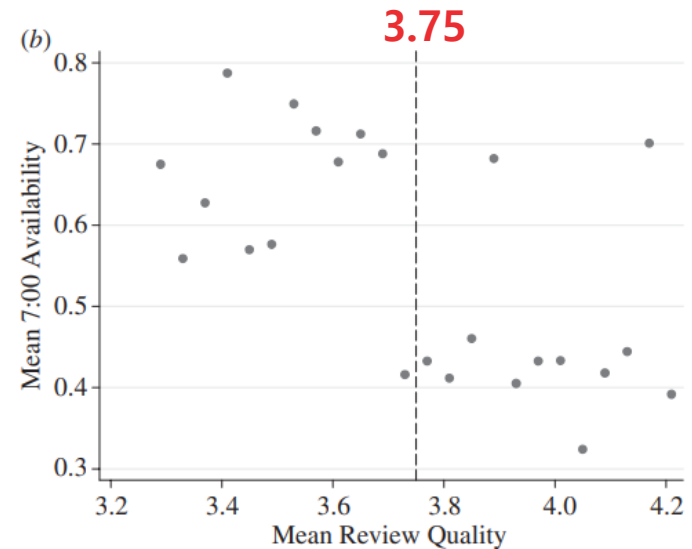
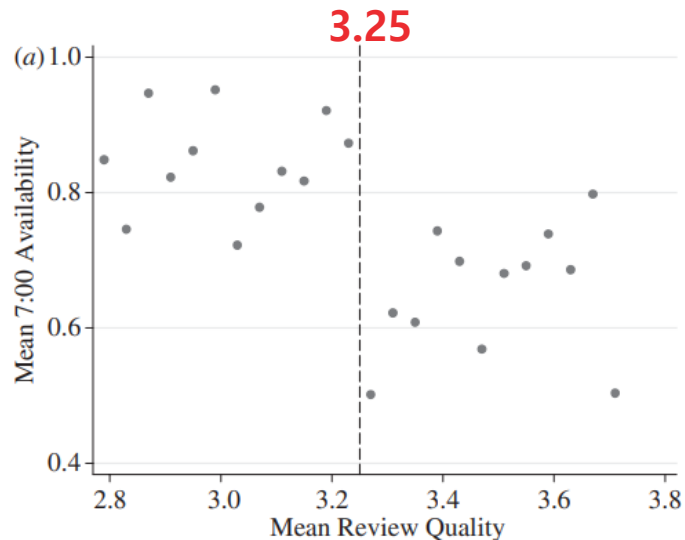
컷오프기반 별점 부여 방식에 따른 효과 분석

- 구체적으로, 해당 연구는 Yelp 평점 표기 방식 상 점수(컷오프 기준으로 0.5단위로 나타내어짐)가 높으면 실제 식당 예약이 많을까? 라는 질문에 답하고자 하였음
 - 그런데 만약 이러한 0.5 단위의 표기 방식이 실제 예약에 영향을 미친다면, 업주들은 컷오프 근처에 있는 자기 업장의 평점을 올리기 위해 평점 조작을 해야 하는 압박을 느낄 수도 있지 않을까?
 - 예) 여러분이 새롭게 식당을 창업했는데 평균 평점이 4.24라서 별이 4개 라면...?

Introduction

별점 부여 방식에 따른 효과

- 해당 연구에서는 각 레스토랑의 저녁 7시 예약 가능 여부를 조사한 후, 주요 반올림 컷오프(3.25, 3.75 등)를 기준으로 아래와 같은 그래프를 제시하였음



위 그래프에 회귀선(regression line)을 그린다고 생각해보자


Introduction


컷오프기반 별점 부여 방식에 따른 효과 분석

- 해당 연구에 따르면 **0.5 별점이 추가될수록 저녁 예약을 최대 약 19%p (49%) 증가시키는 것으로 나타났음**
 - 부가 정보가 없을수록(즉 0.5단위 별점의 영향이 강할수록) 그 영향력이 더 강했음
 - 소비자들이 식당 선택에 있어 **Yelp 평점에 강하게 영향을 받는다**는 점을 시사
- 해당 연구에서는 이와 같은 별점 표기에 따른 효과가 존재한다면 **컷오프 근처 평점을 가진 업주가 별점을 조작할 가능성이 있다고 우려**하였음
 - 우리 식당 리뷰가 4.24점이라면, 가짜 리뷰를 남겨 4.25점을 넘기고 싶은 강력한 유혹을 느끼게 될 수 있음
 - 다만 정량분석 결과 의심이 될만한 패턴은 관찰되지 않았음

Introduction

해당 연구 때문인지는 모르겠지만...





[Yelp for Business](#) [Write a Review](#) [Log In](#) [Sign Up](#)

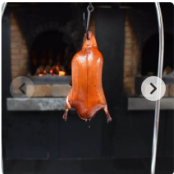
[Restaurants](#) [Home Services](#) [Auto Services](#) [More](#)

[Yelp](#) [Restaurants](#) [Peking Duck](#)


Top 10 Best peking duck Near San Francisco, California [Sort: Recommended](#)

[All](#) [Price](#) [Open Now](#) [Reservations](#) [Offers Online Waitlist](#) [Offers Delivery](#) [Offers Takeout](#)

Sponsored Results






Z & Y Peking Duck

 3.9 (179 reviews)


[Asian Fusion](#) [Chinese](#) [Wine Bars](#) [Chinatown](#)

Closed until 오전 11:30 tomorrow


 Family-owned & operated •  Private events

 Make an Online Reservation


“Very contemporary and modern. There's a pretty extensive menu and pretty diverse clientele. I'd definitely recommend making a reservation. We ordered some beers (\$5 each) and the...” [more](#)

 Delivery

[Find a Table](#)






Kingdom Of Dumplings

 3.7 (2.3k reviews)

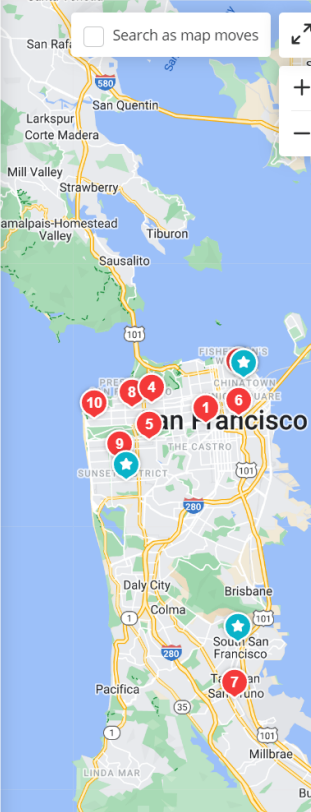
[Dim Sum](#) [Parkside](#)

Closed until 오전 11:00

“Without a doubt a hidden gem. Small, family run, lost place where you can have the best dumplings in town. We'll start a periodical pilgrimage to Parkside to eat their xiao long bao.” [more](#)

 Outdoor seating  Delivery  Takeout

[Start Order](#)



12

What is RDD?

Regression Discontinuity Design

임의 규칙과 자연 실험

- 우리가 사는 세상은 **다양한 임의 규칙**들로 이루어져 있음
 - Yelp에서는 **별점을 5점 만점으로 0.5점 단위로 실제 점수를 반올림하여 시각화**하였음 (왜?)
 - 대한민국에서 초등학교는 아동이 **만 6세 된 날**이 속하는 해의 다음 해 **3월 1일**에 입학이 가능함 (왜?)
 - A대학에서는 **학점이 3.8**을 넘으면 대학원 진학 시 장학금을 받음 (왜??)
 - 대한민국에서는 **만 19세가 되는 해의 1월 1일부터** 술을 구입할 수 있음 (왜?????)
- 이러한 임의 규칙들은 **편의성** 때문에 정해진 경우가 많지만, 데이터 분석가 입장에서 이러한 임의 규칙들은 일종의 **자연 실험 환경**을 만들어주기도 함

Regression Discontinuity Design

회귀불연속설계(Regression Discontinuity Design)의 유래 (1)

- RDD는 **Thistlethwaite and Campbell (1960)**에 의해 개발되었음
 - Regression-discontinuity analysis: An alternative to the ex post facto experiment.
- 당시 미국 정부에서는 대학 진학 희망 고등학생들의 **PSAT 시험 결과**에 따라 **National Merit Scholarship (NMS)**이라는 장학금을 수여하였음
 - 해당 장학금은 PSAT 점수가 특정 점수 이상인 경우에만 수여가 되었음
 - 매년 난이도 등을 고려하여 기준 점수가 변경됨
- Thistlewaite and Campbell (1960)에서는 다음과 같은 질문에 관심이 있었음
 - **NMS 장학금의 수여가 무사히 대학을 졸업하는 데 영향을 줄까?**
 - 분석 대상 인과관계는 무엇인가? 처치군과 대조군은 무엇인가?

Regression Discontinuity Design

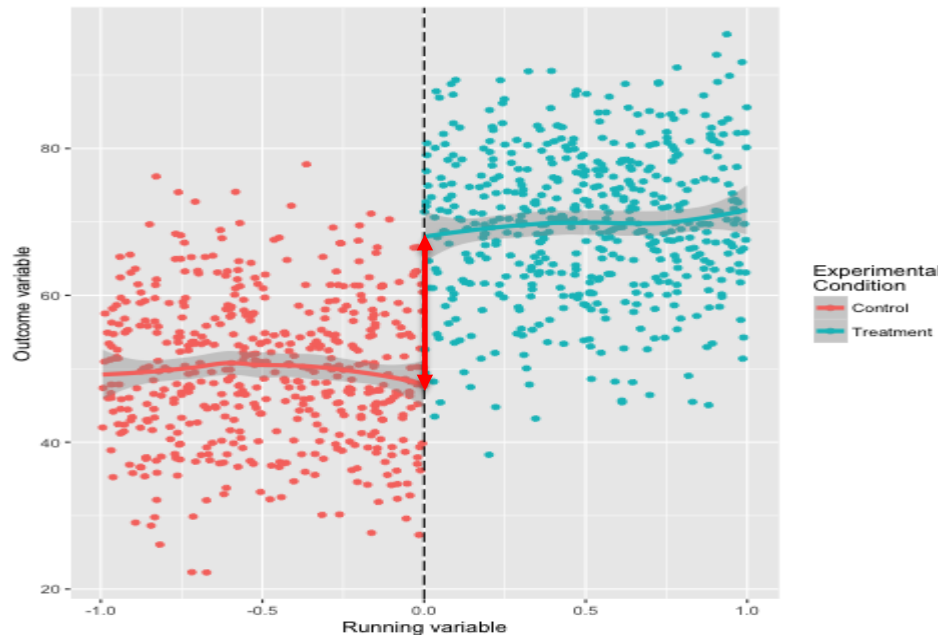
회귀불연속설계(Regression Discontinuity Design)의 유래 (2)

- Naïve한 접근법: **NMS 장학금 수혜자와 미수혜자의 대학 졸업 확률을 비교하면 되는 것 아닌가?**
- 그러나 위와 같은 단순 비교는 장학금 수혜가 대학 졸업에 미치는 인과 관계를 보이기에 적절하지 않음
 - NMS 장학금 수혜자는 **성적이 좋은 학생들**이고, 이들은 대학에 진학해서도 **학업을 우수한 성적으로 이어 나갈 가능성이 높음**
 - 즉, NMS 장학금 수혜자는 아무래도 **대학 진학 후 무사히 졸업할 가능성이 높을 것**
 - 즉, 위와 같은 단순 비교는 **NMS 장학금의 효과를 과대추정할 가능성이 큼**
 - ✓ 장학금을 줘서 졸업한 것이 아니라, 원래 잘하는 학생이라 졸업했을 수도?

Regression Discontinuity Design

회귀불연속설계(Regression Discontinuity Design)의 유래 (3)

- Thistlethwaite and Campbell (1960)에서 제시한 RDD의 핵심은 **컷오프에 인접한 사람들만을 비교**하는 것
 - 처치군: 컷오프를 **살짝 넘어** NMS 장학금을 수혜한 사람
 - 대조군: 컷오프에 **살짝 못미쳐** NMS 장학금을 수혜받지 못한 사람
 - 이 두 그룹의 회귀선에 **불연속(단절)**이 존재하는가?



예) PSAT 1500점이
컷오프라면

1499 이상 1500 미만
vs
1500 이상 1501 미만

비교

Regression Discontinuity Design

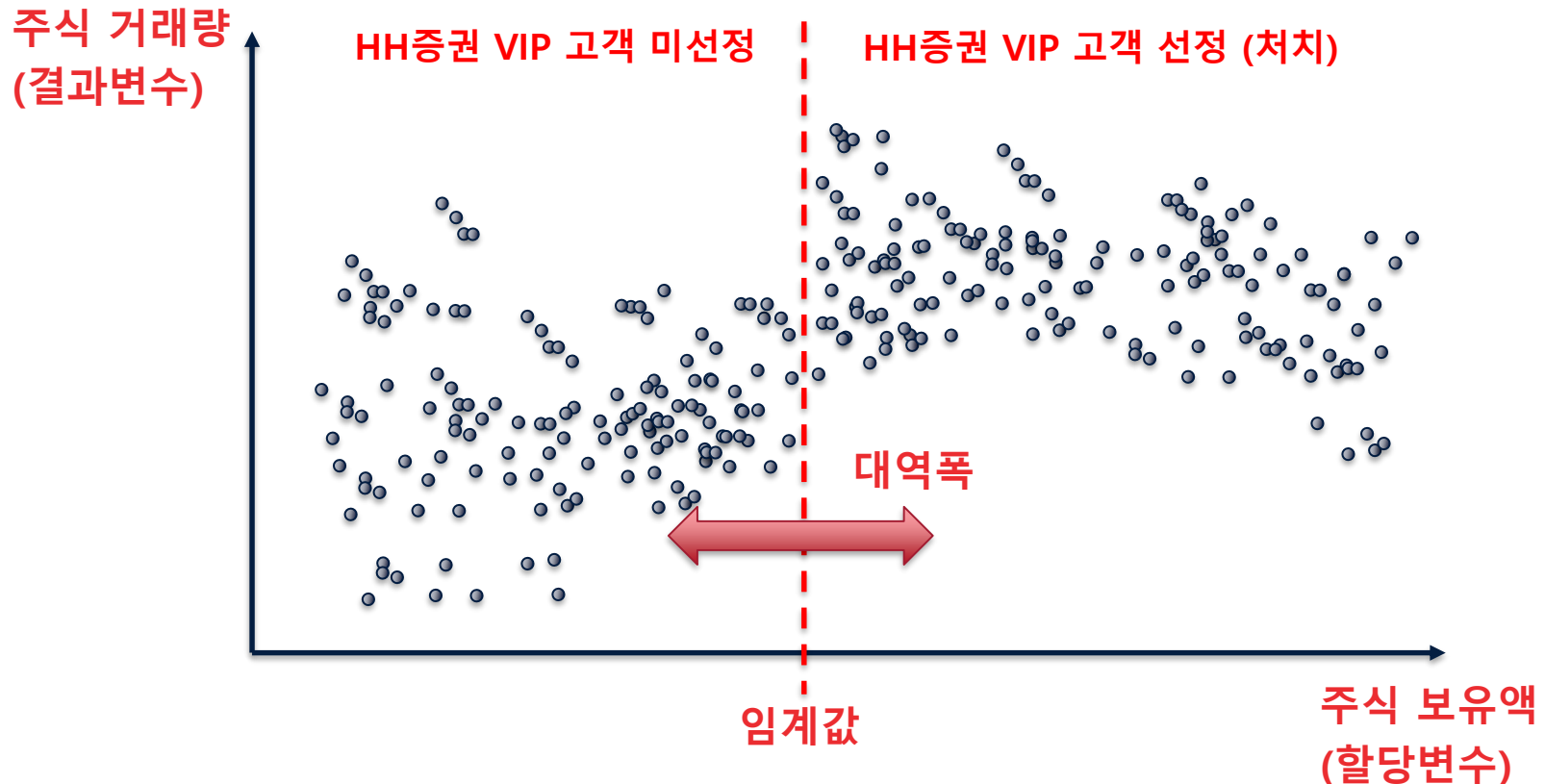
RDD의 주요 개념

- 임계값(cutoff, cutoff)
 - 처치를 받게 되는 조건을 결정하는 기준점
 - ✓ 예) Yelp 평점에서 3.25, 3.75, 4.25, 4.75 등, 장학금 지원 기준 점수
 - 임계값을 기준으로 처치군과 대조군이 구분됨
- 할당 변수(assignment variable, running variable)
 - 임계값 설정의 대상이 되는 변수
 - ✓ 예) Yelp 평점, PSAT 점수 등
 - cf) 결과 변수(outcome variable): 처치에 따라 영향을 받을 것으로 기대하는 변수
 - ✓ 예) 식당 예약률, 대학 졸업률 등
- 대역폭(bandwidth)
 - 배정변수의 특정 임계값 기준으로 어느 범위의 데이터를 포함할 것인지를 나타내는 기준
 - ✓ 예) Yelp 평점 ± 0.25 점, PSAT 점수 ± 1 점 등

Regression Discontinuity Design

RDD 주요 개념 예 HH증권에서는 주식 보유액 일정금액 이상을 기준으로 선정된 고객을 대상으로 VIP 고객 선정 이벤트를 진행하여 각종 혜택을 제공함 (환전 및 거래 수수료 등)

Q. VIP 고객 선정으로 인한 효과(주식 거래량 증가)는 어떻게 측정해야 할까?



Regression Discontinuity Design

인과성 추론에서 RDD의 의의

▪ 간단하면서도 강력한 시각화

- RDD는 인과성 추론 및 주장에 도움이 되는 **강력한 시각화 자료**를 제시할 수 있음 (cf. DID)
- 최초로 제안된 당시(1960)보다도 최근 다양한 분야에서 더 널리 활용됨

▪ 선택 편향에 대응

- **임계값 근처의 관측치들은 유사한 특성을 가진다고 합리적으로 가정**할 수 있는 경우가 많기 때문에 선택 편향에 어느 정도 대응할 수 있음

▪ 준실험 방법론

- 무작위통제실험이 불가능한 상황에서 **적절한 할당변수, 임계값 및 대역폭을 설정**하여 실험과 유사한 분석 환경을 제공하고 강력한 인과성 추론의 근거를 제시할 수 있음

Regression Discontinuity Design

RDD를 활용한 연구들 (1) 특정한 임계값에 의한 처치가 들어가는 경우 유용

기초노령연금이 수급가구의 소득과 소비에 미친 영향: 회귀불연속설계 접근

이석민, 장효진 - 국정관리연구, 2015 - kiss.kstudy.com

... 그래서 본 연구는 준실험방법 중 하나인 회귀불연속설계의 모수추정과 비모수추정을 실시하고 그 결과를 비교하였다. 정책효과를 추정하기 위한 소득과 소비의 분석결과, 기초노령연금의 ...

☆ Save Cite Cited by 7 Related articles All 3 versions »

할당 변수: 소득

처치: 기초노령연금 수급

결과 변수: 소득 및 소비

할당 변수: 저소득학생 수

처치: 교육복지학교 지정

**결과 변수: 학습부진아 비율
(학력 격차 완화)**

교육복지 학교 지정이 학교 간 재정의 수직적 형평성 및 학력격차 완화에 미치는 영향: 회귀불연속 설계를 활용한 인과관계 분석

김경년, 박정신 - 교육행정학연구, 2014 - dbpia.co.kr

... 본 연구는 선행연구에서 제시한 교육특 연구 방법에 대해서 제기한 문제점을 반영하여 준실험적인 연구설계 중 회귀불연속 설계를 활용하여 학력격차에 미치는 영향을 분석하였다. 분석결과...

☆ Save Cite Cited by 8 Related articles

[PDF] 한국프로농구 경기 중 열세 상황에 대한 인식이 경기결과에 미치는 영향: 전망 이론에 근거한 실증연구

김필수, 최준규 - 한국체육학회지 제, 2023 - researchgate.net

... 표본의 경기데이터를 대상으로 회귀불연속설계를 이용하여 전반전 및 3쿼터 종료 시점에서 점수 차에 따른 홈팀 승률의 변화를 분석하였다. 연구결과는 다음과 같다. 첫째, 전반전 종료 후 ...

☆ Save Cite Related articles All 3 versions »

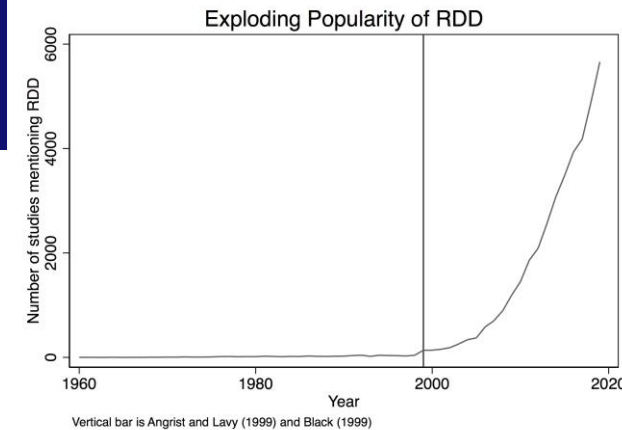
할당 변수: 전반 및 3쿼터 종료 시점의 점수차

처치: 열세 상황 인식

결과 변수: 승률

Regression Discontinuity

RDD를 활용한 연구들 (2)



Using Maimonides' rule to estimate the effect of class size on scholastic achievement

[JD Angrist](#), [V Lavy](#) - The Quarterly journal of economics, 1999 - academic.oup.com

The twelfth century rabbinic scholar Maimonides proposed a maximum class size of 40.

This same maximum induces a nonlinear and nonmonotonic relationship between grade ...

☆ Save Cite Cited by 2833 Related articles All 31 versions Web of Science: 766

할당 변수: 반 인원

처치: 소/대형 반 (Maimonides' rule)

결과 변수: 학습 성과

할당 변수: 거주지

처치: 고급 초등학교

결과 변수: 부동산 가격

Do better **schools** matter? Parental valuation of elementary education

[SE Black](#) - The quarterly journal of economics, 1999 - academic.oup.com

... The evaluation of numerous **school** reforms requires an understanding of the value of better **schools**. Given the difficulty of calculating the relationship between **school** quality and ...

☆ Save Cite Cited by 2524 Related articles All 23 versions Web of Science: 780

Does air quality matter? Evidence from the housing market

[KY Chay](#), [M Greenstone](#) - Journal of political Economy, 2005 - journals.uchicago.edu

We exploit the structure of the Clean **Air** Act to provide new evidence on the capitalization of total suspended particulates (TSPs) **air** pollution into housing values. This legislation ...

☆ Save Cite Cited by 1645 Related articles All 23 versions Web of Science: 550

할당 변수: 대기질 수준

처치: Clean Air Act 적용

결과 변수: 부동산 가격

Regression Discontinuity Design

번외) Maimonides's rule

- 12세기 유대인 철학자인 **Moses Maimonides**가 주장한 교육 원칙
- 이스라엘 공립학교의 반 정원은 40을 초과하면 안된다는 원칙
- 반 정원이 40명을 초과할 경우 반드시 새로운 반이 개설되어야 한다고 주장
 - 40명이라는 명확한 임계값 제시



Regression Discontinuity Design

RDD의 주요 가정

- **할당 변수와 결과 변수의 연속성**
 - 임계값 기준으로 할당 변수와 결과 변수가 자연스러운 연속적인 분포를 가져야 한다는 가정
 - 이러한 연속성을 바탕으로 임계값 주변의 관측치들이 동질적이라는 가정을 할 수 있으며, 이를 바탕으로 처치군과 대조군의 차이가 처치에 의한 것이라고 주장할 수 있음
- **조작 가능성의 부재**
 - 관측치들이 처치를 받거나 받지 않기 위해 할당 변수를 조작할 수 없어야 하며, 임계값 주변에서 분포가 자연스럽게 형성되어야 함
- **임계값 주변의 동질성**
 - 임계값 근처의 관측치들이 비교적 동질적이라는 가정으로, 다른 관찰되지 않은 특성들 역시 유사할 것이라고 가정

Regression Discontinuity Design

RDD의 한계 (및 이에 따른 주의사항) (1)

■ 일반화의 어려움

- 할당 변수 임계값 근처의 관측치들만이 분석에 포함됨
- 이에 따라 일부 관측치들이 분석에서 제외될 수 있으며, 임계값 근처의 관측치들이 관심 대상 전체 모집단을 대표하지 않는다면 일반화 범위가 제한됨 (cf. PSM)

■ 대역폭 선택의 어려움

- 임계값으로부터 얼마나 떨어진 관측치까지 분석에 포함하느냐는 매우 중요한 문제이지만 적절히 결정하기 어려운 경우도 많음
- 너무 좁은 대역폭을 선택하면 관측치 수가 줄어들어 분석의 정밀도가 떨어질 수 있으며, 너무 넓은 대역폭을 선택하면 처치군과 대조군 사이의 동질성이 약해지고 임계값 근처의 불연속성을 제대로 포착하지 못할 수 있음

Regression Discontinuity Design

RDD의 한계 (및 이에 따른 주의사항) (2)

■ 데이터 제한

- 임계값 근처의 관측치가 충분하지 **않다면** 적절한 분석이 어려울 수 있음
- 예를 들어, 적절한 수준의 관측치가 확보되지 않는다면 **적절한 통계적 검정을 수행하기 어렵고** 분석 결과의 신뢰도가 저하됨

■ 조작 가능성

- 할당 변수에 조작 가능성이 있는 경우 **임계값 근처 관측치의 분포를 왜곡**시킬 수 있기 때문에 분석의 유효성에 영향을 미침 (RDD 주요 가정 위배)
- 예) 장학금을 받기 위한 임계값을 학생들이 사전에 알고 있다면, 임계값 인근의 학생들이 해당 점수를 얻기 위한 별도의 노력을 하는 경우가 있을 수 있음 → 분포가 어떻게 변화할까?

Conducting RDD Analysis

Conducting RDD Analysis

RDD 분석의 주요 단계

- RDD 분석의 주요 단계는 다음과 같이 설정할 수 있음
 1. 할당변수 및 임계값 결정
 2. 대역폭 및 모형 선택
 3. RDD 추정 및 결과 해석

Conducting RDD Analysis

1. 할당변수 및 임계값 설정

- RDD를 활용한 연구 설계에 해당하는 부분
- 할당변수: 처치의 기준이 되는 변수
 - 인과성 분석: 처치 → 결과 변수
- 임계값은 처치 여부를 결정하는 할당변수의 기준값으로, 다양한 배경에서 설정될 수 있음
 - 법/정책: 음주 가능 연령, 연 매출 30억 이상 사업장 지역화폐 사용 제한 등
 - 기타 (임의) 규칙: Yelp의 평점 반올림 방식, A대학 합격 점수, B기업의 우수고객 선정 방식, Maimonide's rule 등

Conducting RDD Analysis

분석 대상 데이터: 미국 MLDA 데이터

- MLDA(minimum legal drinking age): 최소 합법 음주 연령
 - 미국의 경우 만 21세
 - 우리나라의 경우 만 19세가 되는 해의 1월 1일
 - 단, 음주 행위 자체보다는 주류 구매 및 주점 출입에 대한 제한이 존재



Conducting RDD Analysis

분석 대상 데이터: 미국 MLDA 데이터

- 음주가 법적으로 허용된다고 부상 및 사망이 늘어나나? (효과가 금방 나타나나?)
 - 음주가 법적으로 허용되어서 **술을 더** 마시게 됨
 - 음주로 인해 건강이 악화되거나 사망하는 것 외에도 다양한 부상 및 사망 요인들이 존재
 - 교통사고(음주운전 등), 폭행시비, 정신건강 악화 등
- 참고 연구:
 - Carpenter, C., & Dobkin, C. (2009). The effect of alcohol consumption on mortality: regression discontinuity evidence from the minimum drinking age. *American Economic Journal: Applied Economics*, 1(1), 164-182.

Conducting RDD Analysis

미국 MLDA 데이터 변수 둘러보기

- 변수 목록
 - **age**: 나이 (개월 기준 소수점 표기)
 - **limit**: 음주 제한 연령 (모두 21로 동일)
 - **age_d**: 나이 - 음주 제한 연령
 - **treat**: 음주 합법 허용 (21세 이상) 여부
 - **visit**: 병원 응급실 방문 횟수

Conducting RDD Analysis

MLDA 연구 질문과 변수

- 분석 대상 인과 **treat** → **visit**
- 음주가 법적으로 허용되면 부상 및 사망이 많아질까?
 - 처치: 음주 법적 허용 (**treat**)
 - ✓ 할당 변수: 나이 (**age**)
 - ✓ 임계값: 만 21세 (실제 정책에 의한 기준) (**limit**)
 - ✓ 대역폭: ?? (추후 2년으로 일단 설정)
 - 결과변수: 응급실 방문 횟수 (**visit**)
 - 처치군: 만 21세 이상, 음주가 허용됨
 - 대조군: 만 21세 미만, 음주가 허용되지 않음
- RDD 분석 상황에서는 할당변수를 알면 처치 여부를 알 수 있으며, 처치 여부는 할당 변수의 불연속 함수임

Conducting RDD Analysis

실습: 기본 시각화

#데이터 활용 패키지 및 시각화 패키지 설치 및 불러오기

```
install.packages("dplyr")  
install.packages("ggplot2")  
library(dplyr)  
library(ggplot2)
```

#데이터 불러오기

```
mydata <- read.csv("BEDA_4_RDD_data_carpenterdobkin.csv", sep=',')
```

#데이터 둘러보기

```
head(mydata)  
summary(mydata)
```

#시각화 기본

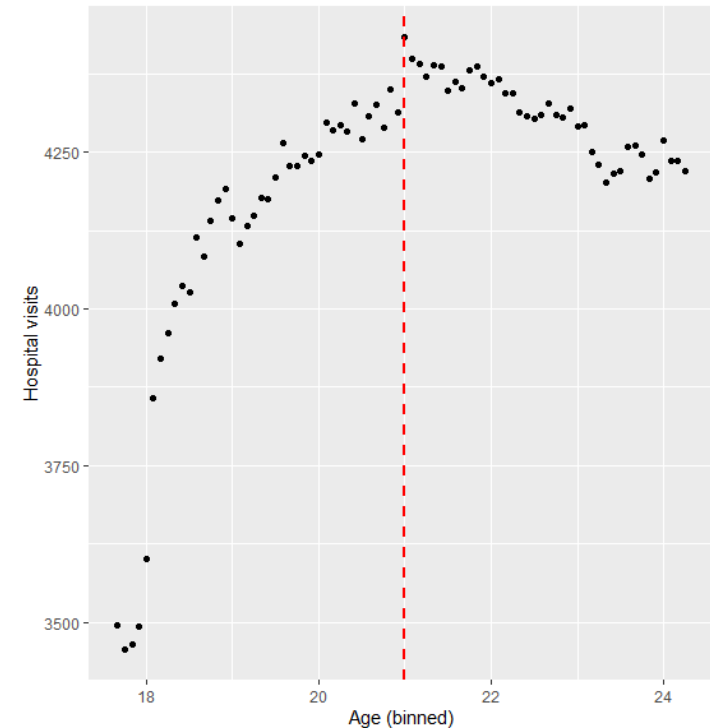
```
mydata %>% x축, y축 칼럼명
```

```
ggplot(aes(x = age, y = visit)) +  
geom_point() +
```

컷오프 선 그어주기 (위치, 색, 굵기, 선 타입 지정 가능)

```
geom_vline(xintercept = 21, color = "red", size = 1, linetype = "dashed") +  
labs(y = "Hospital visits", x = "Age (binned)")
```

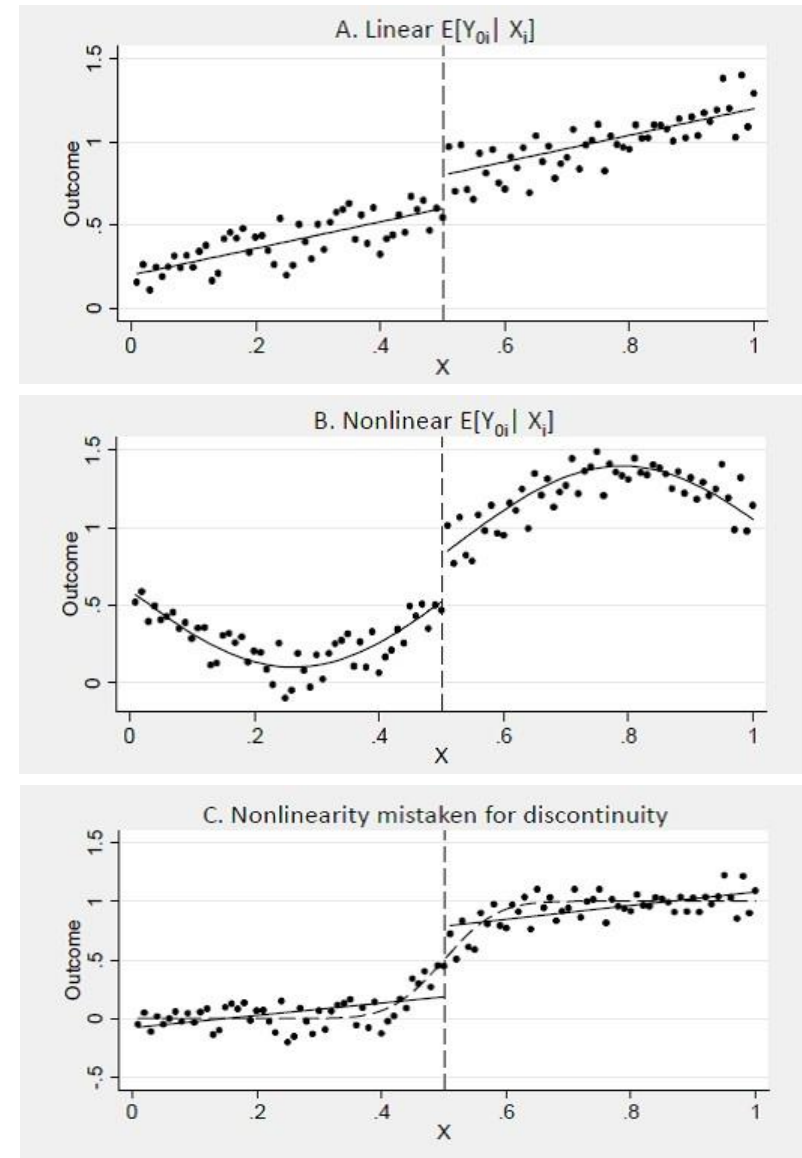
X,Y 라벨



Conducting RDD Analysis

2. 대역폭 및 모형 선택 - 단절의 식별

- RDD에서는 할당변수의 컷오프를 기준으로 나타나는 처치군과 대조군 사이의 단절(discontinuity)을 식별하고자 함
 - [A] 뚜렷한 선형 단절
 - [B] 뚜렷한 비선형 단절
 - [C] 비선형 관계를 단절로 오인?
- C와 같은 사례에서 어디부터 단절로 인정하고 어디부터 그렇게 하지 않을지를 정하는 것은 어려운 문제
 - 어느 정도 연구자의 자의성이 존재
 - 그렇기 때문에 RDD를 활용할 경우 연구자는 여러 대역폭 및 모형 선택 결과를 함께 보고하는 것이 바람직



Conducting RDD Analysis

2. 대역폭 및 모형 선택 - 대역폭

- RDD의 세부 사항이라고 할 수 있는 대역폭 및 모형을 선택하는 단계
- 대역폭은 임계값을 기준으로 데이터의 포함 범위를 나타냄
 - RDD의 핵심 아이디어 중 하나는 컷오프 근처 관측치에 집중하는 것
 - **대역폭이 좁은 경우:** 처치군과 대조군의 동질성이 강할 가능성이 높으나, 관측치 수가 부족하여 통계적 검정이 어려워지거나 결과의 일반화가 비교적 어려울 수 있음
 - **대역폭이 넓은 경우:** 관측치 수가 풍부하고 결과의 일반화가 비교적 용이할 수 있으나, 처치군과 대조군의 동질성이 약해져 처치로 인한 효과를 적절히 포착해내지 못할 수 있음
 - ✓ 예) MLDA의 대역폭을 20년으로 늘리면?
 - 최적의 대역폭 수준을 어떻게 알 수 있는가? → **모름**
 - ✓ 물론 의사결정을 지원할 수 있는 지표 및 기준이 존재하기는 함
 - 따라서, **여러 대역폭을 수준을 시도해보고 이를 비교하여 최종 대역폭을 결정**하는 방식이 권장됨
 - ✓ 해당 결과도 보고하는 것이 바람직함

Conducting RDD Analysis

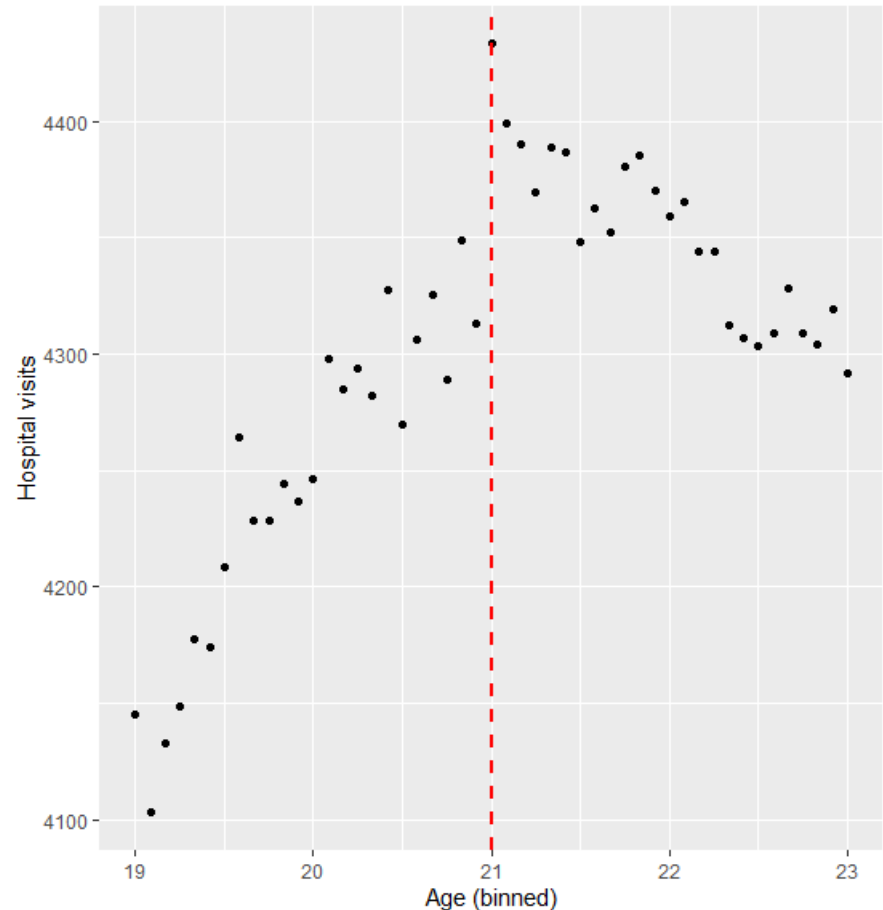
실습: 대역폭을 2년(24개월)으로 설정 후 다시 그림을 그려보자

#대역폭 2년 설정

```
filtered_data <- mydata %>%  
  filter(age >= 19 & age <= 23)
```

#19~23세 시각화

```
filtered_data %>%  
  ggplot(aes(x = age, y = visit)) +  
  geom_point() +  
  geom_vline(xintercept = 21, color = "red", size = 1, linetype = "dashed") +  
  labs(y = "Hospital visits", x = "Age (binned)")
```



Conducting RDD Analysis

2. 대역폭 및 모형 선택 - 모형

본 수업은 모수적 방식에 집중

- RDD의 모형 설정은 크게 두 가지 방식이 활용될 수 있음

- 모수적(parametric) 방식

- ✓ 처치군과 대조군에 대해 특정 함수 형태의 모형을 가정하고 그 모수(parameter)를 추정하는 방식

- 예) 선형회귀모형

- ✓ 특정 함수 형태의 모형을 기반으로 결과의 해석이 용이하다는 장점이 있으나, 잘못된 모형 형태를 가정하였을 경우 편의(bias)가 발생할 수 있음

- 그렇다면 어떤 형태의 모형이 가장 적절한가? → 모름

- 따라서 모수적 방식을 활용하는 경우 **다양한 형태의 모형을 시도**하고 모형 형태에 따라 결과가 어떻게 변화하는지를 제시하는 것이 바람직

- 비모수적(nonparametric) 방식

- ✓ 처치군과 대조군에 대해 특정 모형의 형태를 가정하지 않는 방식

- 평균 비교, 커널 밀도 추정, 로컬 회귀 등

Conducting RDD Analysis

3. RDD 추정 및 결과 해석

- RDD에서 주요 결과를 추정하고 해석하는 단계
- 가장 기본적인 형태의 모수적 방식을 살펴보자

$$Y_i = \beta_0 + \beta_1 Treat_i + \beta_2 (X_i - X_c) + \beta_3 Treat_i (X_i - X_c)$$

- Y_i : 관측치 i 의 결과 변수
- $Treat_i$: 관측치 i 의 처치 여부
- X_i : 관측치 i 의 할당변수 값
- X_c : 할당변수의 컷오프 값
- $X_i - X_c$: (나이차) 처치군 여부에 따라 양/음의 값을 가지게 됨
 - ✓ 해당 항을 추가함으로써 선형회귀분석에서의 Y 절편을 X_c (컷오프) 위치로 옮겨올 수 있음
 - ✓ 예: 컷오프가 21이고 관측치의 할당변수값이 22라면, $22-21=1$

Conducting RDD Analysis

3. RDD 추정 및 결과 해석 – 차이값 활용 이유

- X_i 대신 $X_i - X_c$ 를 모형에서 활용하는 이유
 - $X_i - X_c = X_i'$ 이라고 하면:

$$Y_i = \beta_0 + \beta_1 Treat_i + \beta_2 X_i' + \beta_3 Treat_i X_i'$$



Conducting RDD Analysis

3. RDD 추정 및 결과 해석 – 처치군과 대조군의 회귀식

- 처치군과 대조군에 대해 회귀식이 어떻게 달라질까?

$$Y_i = \beta_0 + \beta_1 Treat_i + \beta_2 X_i' + \beta_3 Treat_i X_i'$$

- 만약 위 식에서 $Treat == 0$ 이라면 (대조군)

$$Y_i = \underbrace{\beta_0}_{\text{절편}} + \underbrace{\beta_2 X_i'}_{\text{기울기}}$$

- 반면, 위 식에서 $Treat == 1$ 이라면 (처치군)

$$Y_i = \underbrace{(\beta_0 + \beta_1)}_{\text{절편}} + \underbrace{(\beta_2 + \beta_3) X_i'}_{\text{기울기}}$$

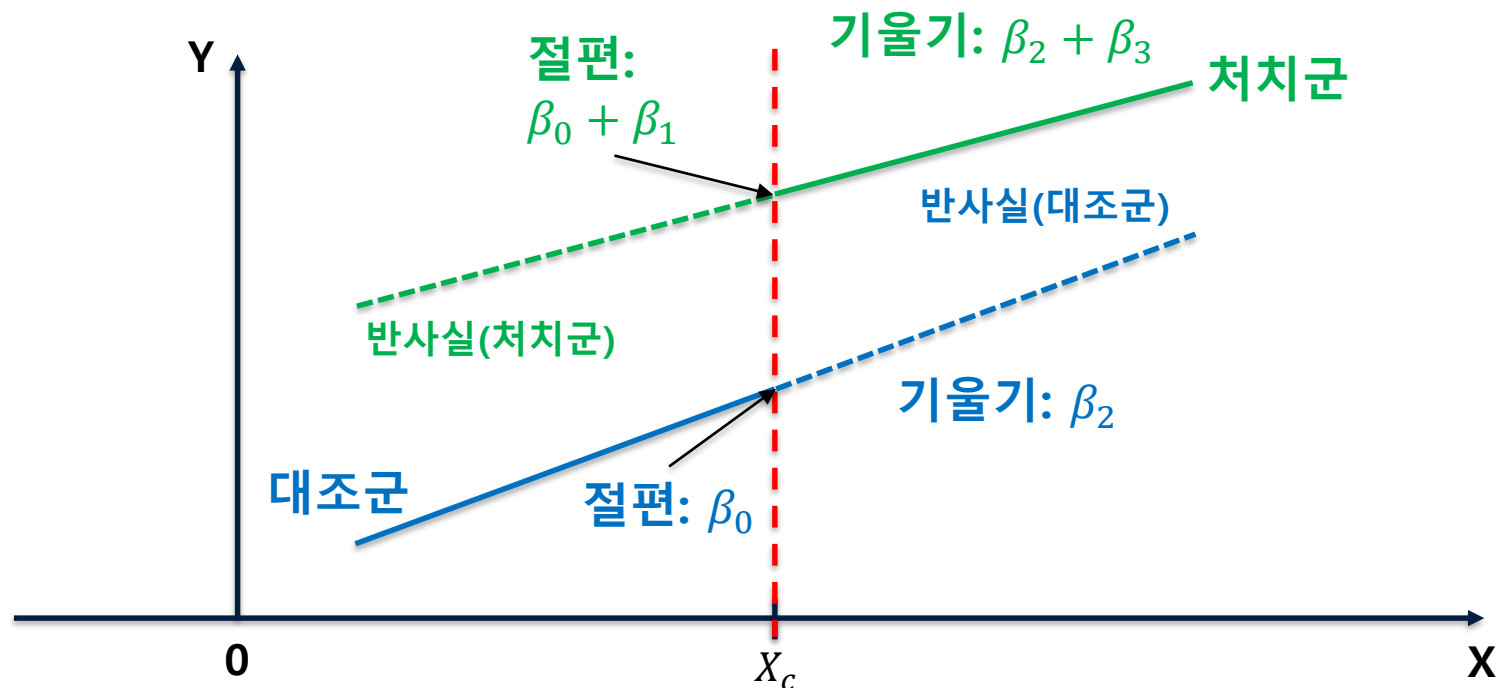
Conducting RDD Analysis

3. RDD 추정 및 결과 해석 - 시각화 (1)

- 즉, 아래 식 추정 결과에 따라....

$$Y_i = \beta_0 + \beta_1 \text{Treat}_i + \beta_2 X_i' + \beta_3 \text{Treat}_i X_i'$$

아래 그림에서
 β_1 과 β_3 의 부호는 무엇일까?

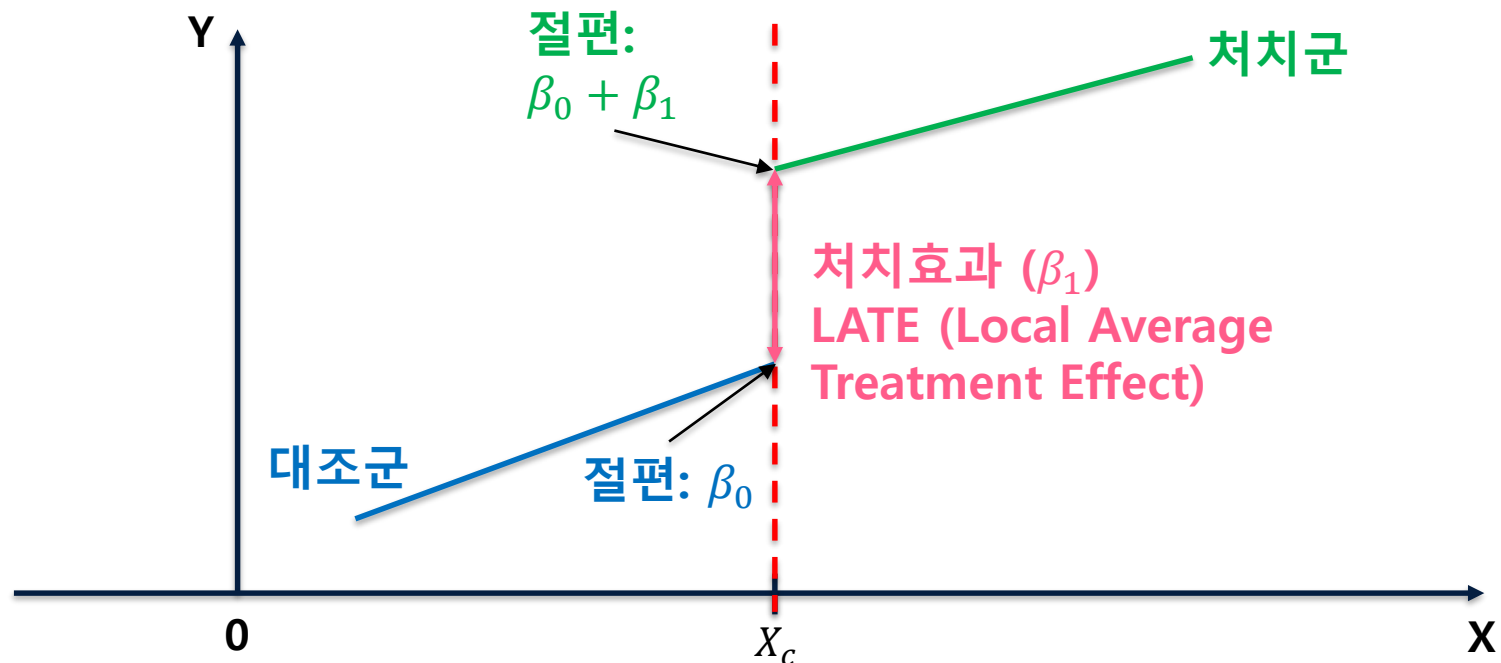


Conducting RDD Analysis

3. RDD 추정 및 결과 해석 - 시각화 (2)

- 즉, 아래 식 추정 결과에 따라....

$$Y_i = \beta_0 + \beta_1 \text{Treat}_i + \beta_2 X_i' + \beta_3 \text{Treat}_i X_i'$$



Conducting RDD Analysis

3. RDD 추정 및 결과 해석 – LATE

- RDD에서는 컷오프를 기준으로 바로 위와 바로 아래에 있는 관측치의 비교를 통해 처치 효과를 추정함
 - 이를 위해 처치군과 대조군의 추세를 선형회귀모형을 이용해 분석하고 컷오프 근처에서의 차이를 비교함
- 이러한 효과는 임계값 근처의 관측치에 집중하므로, 이를 **지역 평균 처리 효과(LATE, Local Average Treatment Effect)**라고 부르기도 함
 - 임계값 근처에서만 국지적으로 유효한 처치 효과임
 - 컷오프 값이 변경되었을 때에도 해당 처치효과가 유지된다고 보기 어려울 수 있음
 - ✓ 예) 합법 음주 허용 컷오프가 90세가 된다면?
- 때로는 LATE와 함께 **처치군과 대조군의 기울기 차이**에 관심이 있는 경우도 있음

Conducting RDD Analysis

실습: 불연속적인 패턴을 고려하지 않은, 전체에 대한 회귀분석 수행 (1)

$$Y_i = \beta_0 + \beta_1 X_i'$$

#단순 선형회귀모형 추정

```
mymodel <- lm(visit ~ age_d, data = filtered_data)
```

```
summary(mymodel)
```

#단순 선형회귀모형 결과 시각화

```
filtered_data %>%
```

```
  ggplot(aes(x = age, y = visit)) +
```

```
  geom_point() +
```

```
  geom_vline(xintercept = 21, color = "red", size = 1, linetype = "dashed") +
```

```
  geom_smooth(method = "lm", color = "blue", se = FALSE) +
```

```
  labs(y = "Hospital visits", x = "Age (binned)")
```

선형회귀모형 결과 추가

lm: 선형회귀모형임을 나타냄

(별도 지시 없을 시 x와 y의 단순선형회귀)

se: 신뢰구간 표시 여부

Conducting RDD Analysis

실습: 불연속적인 패턴을 고려하지 않은, 전체에 대한 회귀분석 수행 (2)

$$Y_i = \beta_0 + \beta_1 X_i'$$

```
Call:
lm(formula = visit ~ age_d, data = filtered_data)
```

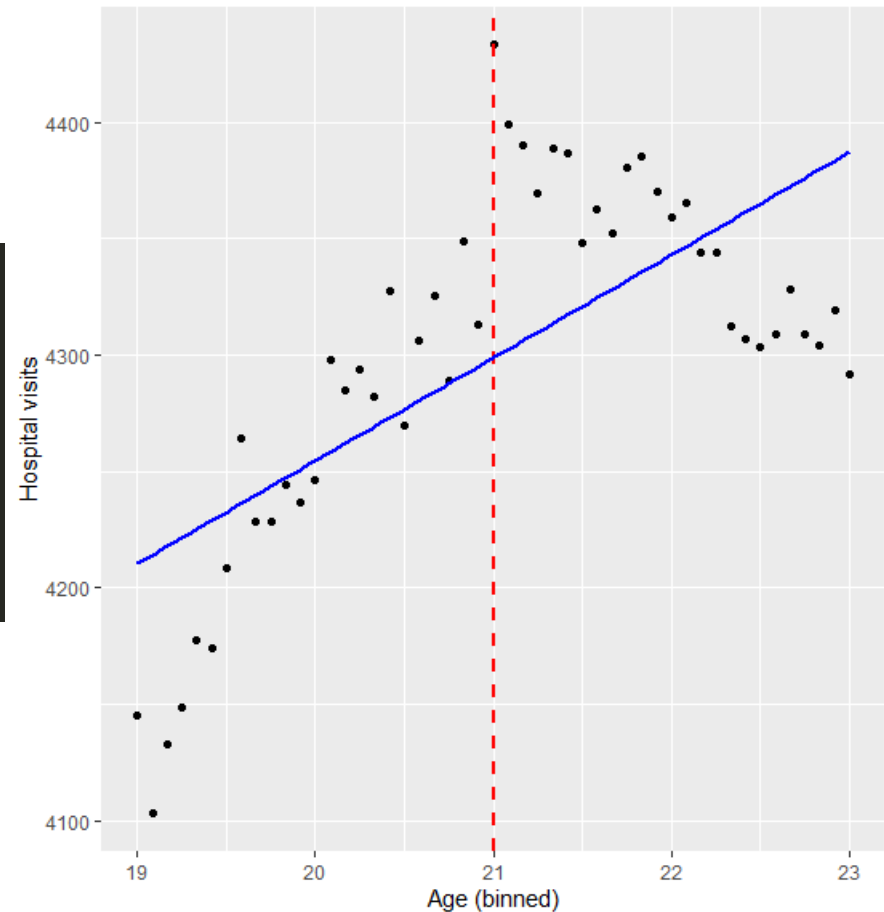
Residuals:

Min	1Q	Median	3Q	Max
-110.835	-47.669	1.191	38.234	134.819

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	4298.687	7.909	543.488	< 2e-16 ***
age_d	44.188	6.711	6.584	3.5e-08 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1



Conducting RDD Analysis

실습: RDD 회귀모형 추정 및 시각화 (1)

$$Y_i = \beta_0 + \beta_1 Treat_i + \beta_2 X_i' + \beta_3 Treat_i X_i'$$

#RDD 회귀모형 추정

```
mymodel2 <- lm(visit ~ treat + age_d + treat:age_d, data = filtered_data)
```

```
summary(mymodel2)
```

#RDD 회귀모형 결과 시각화

```
filtered_data %>%
```

```
ggplot(aes(x = age, y = visit, color = factor(treat))) +
```

```
geom_point() +
```

```
geom_vline(xintercept = 21, color = "red", size = 1, linetype = "dashed") +
```

```
geom_smooth(data = subset(filtered_data, age < 21), method = "lm", color = "cornflowerblue", se = FALSE) +
```

```
geom_smooth(data = subset(filtered_data, age >= 21), method = "lm", color = "forestgreen", se = FALSE) +
```

```
scale_color_manual(values= c("0" = "cornflowerblue", "1" = "forestgreen"))+
```

```
labs(y = "Hospital visits", x = "Age (binned)")
```

색깔을 어떤 기준으로 칠할 것인지 설정 (아래 `scale_color_manual`과 연계)
(factor: 수치 데이터인 `treat`을 범주형 데이터로 인식하게 해 줌)

조건에 따른 회귀선 추가

관측치 포인트도 색칠

Conducting RDD Analysis

실습: RDD 회귀모형 추정 및 시각화 (2)

$$Y_i = \beta_0 + \beta_1 Treat_i + \beta_2 X_i' + \beta_3 Treat_i X_i'$$

Coefficients:

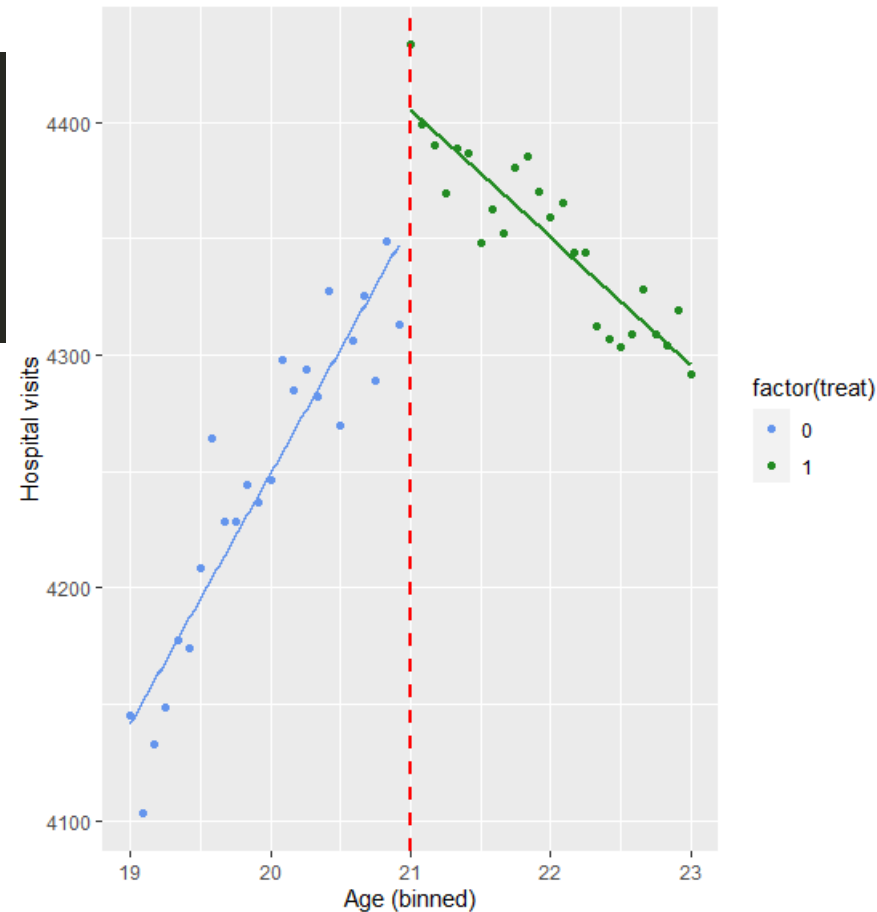
	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	4356.470	9.165	475.323	< 2e-16	***
treat	48.584	12.464	3.898	0.00032	***
age_d	107.260	7.697	13.935	< 2e-16	***
treat:age_d	-161.838	10.567	-15.316	< 2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.'

Q1. 음주 법적 허용의 처치효과는?

Q2. 대조군(처치군)의 21세 컷오프 근처 절편은?

Q3. 대조군(처치군)에서 나이를 한 살 더 먹을 때마다
응급실 방문 횟수 변화량은?



Conducting RDD Analysis

실습: RDD 회귀모형 추정 및 시각화 (3)

■ 모수 추정 결과

$$Y_i = \beta_0 + \beta_1 Treat_i + \beta_2 X_i' + \beta_3 Treat_i X_i'$$

Variables	Coefficient	Std.Error	P-value
Intercept (β_0)	4356.470	9.165	0.000
Treat (β_1)	48.584	12.464	0.000
X_i' (β_2)	107.260	7.697	0.000
$Treat_i: X_i'(\beta_3)$	-161.838	10.567	0.000

- 음주 법적 허용이라는 처치효과(β_1)의 추정치가 유의함
 - ✓ 즉, 21세라는 컷오프 근처에서 음주 법적 허용이라는 처치가 응급실 방문 횟수에 미치는 영향이 유의미함
- 대조군과 처치군의 컷오프 근처 절편과 기울기는?
 - ✓ 대조군: 절편은 4356.5 & 기울기는 107.3
 - ✓ 처치군: 절편은 4405.1 & 기울기는 -54.6

Conducting RDD Analysis

3. RDD 추정 및 결과 해석 – 비선형 추세 존재 가정 시

경우에 따라 고차항 더 추가 가능
3차항, 4차항...

- 만약 비선형적인 추세가 존재한다고 가정한다면?

$$Y = \beta_0 + \beta_1 Treat + \beta_2 X_i' + \beta_3 (X_i')^2 + \beta_4 Treat_i X_i + \beta_5 Treat_i (X_i')^2$$

- 만약 위 식에서 $Treat == 0$ 이라면 (대조군)

$$Y_i = \beta_0 + (\beta_2 + \beta_3 X_i') X_i'$$

- 반면, 위 식에서 $Treat == 1$ 이라면 (처치군)

$$Y_i = (\beta_0 + \beta_1) + (\beta_2 + \beta_3 X_i' + \beta_4 + \beta_5 X_i') X_i'$$

Conducting RDD Analysis

실습: 비선형 RDD 회귀모형 추정 및 시각화 (1)

$$Y = \beta_0 + \beta_1 Treat + \beta_2 X_i' + \beta_3 (X_i')^2 + \beta_4 Treat_i X_i' + \beta_5 Treat_i (X_i')^2$$

#RDD 비선형 회귀모형 추정

```
Mymodel3 <- lm(visit ~ treat + age_d + I(age_d^2) + treat:age_d + treat:I(age_d^2), data = filtered_data)
```

```
summary(mymodel3)
```

제공항 표현 방법:

알파벳 대문자 I(아이) 안에 넣고 몇제곱인지 ^ 활용하여 표기

#RDD 비선형 회귀모형 결과 시각화

```
filtered_data %>%
```

```
  ggplot(aes(x = age, y = visit, color = factor(treat))) +  
    geom_point() +  
    geom_vline(xintercept = 21, color = "red", size = 1, linetype = "dashed") +  
    geom_smooth(data = subset(filtered_data, age < 21), method = "lm", formula = y ~ x + I(x^2),  
color = "cornflowerblue", se = FALSE) +  
    geom_smooth(data = subset(filtered_data, age >= 21), method = "lm", formula = y ~ x + I(x^2),  
color = "forestgreen", se = FALSE) +  
    scale_color_manual(values= c("0" = "cornflowerblue", "1" = "forestgreen"))+  
    labs(y = "Hospital visits", x = "Age (binned)")
```

제공항 포함 분석임을 나타냄

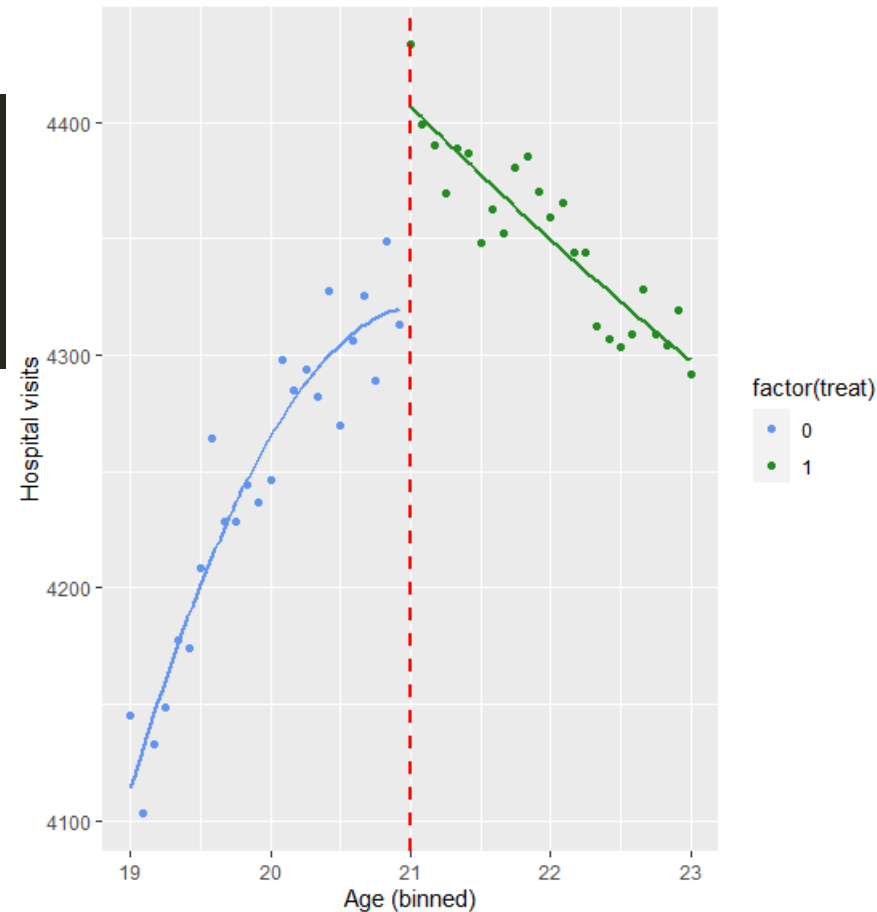
Conducting RDD Analysis

실습: 비선형 RDD 회귀모형 추정 및 시각화 (2)

$$Y = \beta_0 + \beta_1 Treat + \beta_2 X_i' + \beta_3 (X_i')^2 + \beta_4 Treat_i X_i + \beta_5 Treat_i (X_i')^2$$

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	4320.560	13.053	331.008	< 2e-16	***
treat	86.095	16.978	5.071	8.02e-06	***
age_d	7.817	28.869	0.271	0.787845	
I(age_d^2)	-47.732	13.453	-3.548	0.000953	***
treat:age_d	-67.408	38.283	-1.761	0.085389	.
treat:I(age_d^2)	50.239	18.123	2.772	0.008198	**



- Q1. 음주 법적 허용의 처치효과는?
- Q2. 대조군(처치군)의 21세 절편은?
- Q3. 대조군(처치군)에서 나이를 한 살 더 먹을 때마다
응급실 방문 횟수 변화량은?

(모든 모수가 통계적으로 유의하다고 치고 한번 계산해보자)

Conducting RDD Analysis

실습: 비선형 RDD 회귀모형 추정 및 시각화 (3)

$$Y = \beta_0 + \beta_1 Treat + \beta_2 X_i' + \beta_3 (X_i')^2 + \beta_4 Treat_i X_i + \beta_5 Treat_i (X_i')^2$$

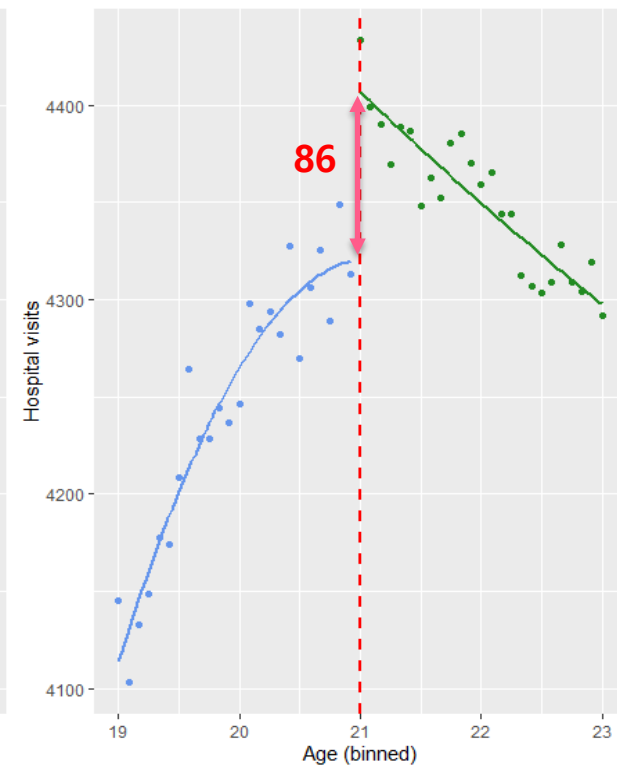
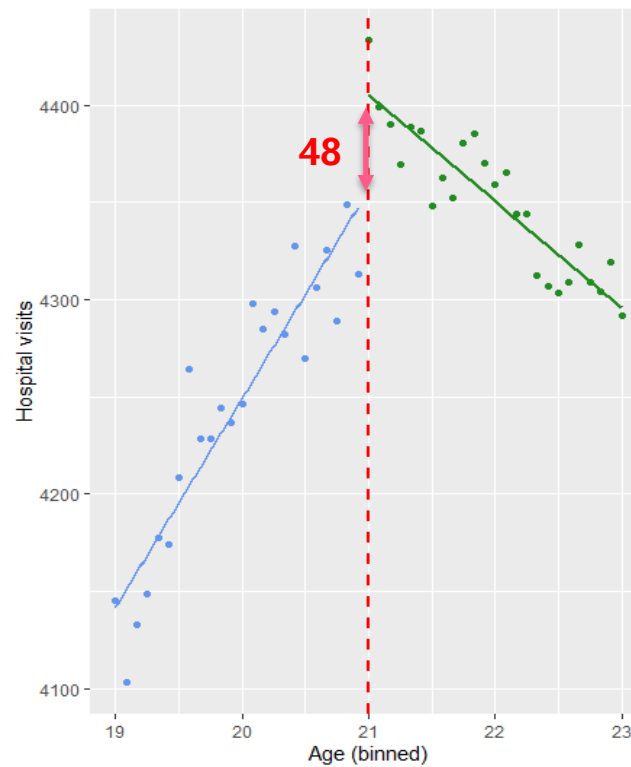
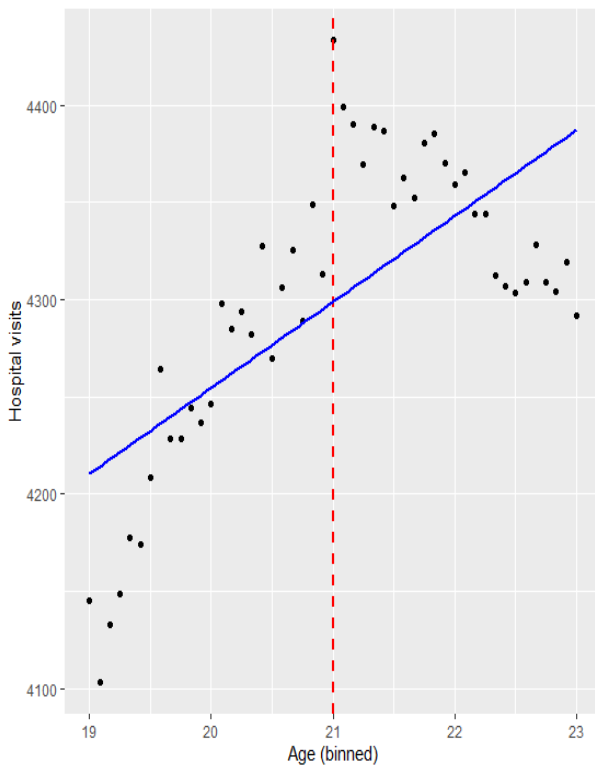
■ 모수 추정 결과

Variables	Coefficient	Std.Error	P-value
Intercept (β_0)	4320.560	13.053	0.000
Treat (β_1)	86.095	16.978	0.000
X_i' (β_2)	7.817	28.869	0.788
$(X_i')^2$ (β_3)	-47.732	13.453	0.001
$Treat_i: X_i'$ (β_4)	-67.408	38.283	0.085
$Treat_i: (X_i')^2$ (β_5)	50.239	18.123	0.008

- 음주 법적 허용이라는 처치효과(β_1)의 추정치가 유의함
 - ✓ 이번에는 86으로 선형인 경우보다 차이가 더 크게 나타남
- 대조군과 처치군의 컷오프 근처 절편과 기울기는?
 - ✓ 대조군: 절편은 4320.6 & 기울기는 $7.8 - 47.7 X_i'$
 - ✓ 처치군: 절편은 4406.7 & 기울기는 $-59.6 + 2.5 X_i'$

Conducting RDD Analysis

실습: 시각화 결과 모아보기



Conducting RDD Analysis

도전 과제

- **다른 모형 명세**로도 분석해보자
 - 3차항 포함 비선형 모형 등
 - 결과가 크게 차이 나는가?
- **대역폭**을 다르게 하여 분석해보자
 - 18개월 등
 - 결과가 크게 차이 나는가?

Conducting RDD Analysis

보너스) 도입 사례의 Yelp 분석 결과를 이제 세부적으로 살펴보자

- 할당변수: Yelp 별점
- 컷오프: 3.25, 3.75, 4.25 (해당 컷오프를 넘으면 각각 3.5, 4, 4.5 별점 제시됨)
- 결과변수: 저녁 6, 7, 8시 예약 가능 여부

Regression Discontinuity Results at Individual Thresholds

Yelp display rating	6:00 PM availability			7:00 PM availability			8:00 PM availability		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
3.5 Yelp stars	-0.079 (0.086)			-0.213** (0.096)			-0.150* (0.080)		
4 Yelp stars		-0.101 (0.075)			-0.192** (0.093)			-0.095 (0.086)	
4.5 Yelp stars			0.004 (0.185)			-0.113 (0.127)			-0.119 (0.149)
Yelp rating	-0.228 (0.201)	0.145 (0.203)	-0.131 (0.230)	0.082 (0.216)	0.024 (0.255)	-0.022 (0.271)	0.088 (0.180)	0.008 (0.218)	-0.321 (0.276)
Yelp rating × Yelp star	0.372 (0.287)	-0.275 (0.309)	-2.934** (1.342)	-0.057 (0.335)	-0.048 (0.375)	-1.817*** (0.674)	-0.080 (0.282)	-0.329 (0.352)	-1.324 (0.869)
Observations	8,705	11,858	5,597	8,705	11,858	5,597	8,705	11,858	5,597

Notes. Contains RD estimates of the effects of an additional Yelp half-star on availability. Availability measures indicate whether the reservations were available at that time on Thursday, Friday or Saturday when queried 36 hours in advance. Standard errors are clustered at the restaurant level. Asterisks denote significance levels at: *10%, **5% and ***1%.

(5)
표시되는 별점이
4점인지 여부에 따라
(3.75 근처)
예약가능 여부
19.2%p 변화

Conducting RDD Analysis

보너스) 도입 사례의 Yelp 분석 결과를 이제 세부적으로 살펴보자

Regression Discontinuity Results at Pooled Thresholds

	(1)	(2)	(3)	(4)	(5)
<i>Panel (a): 6:00 PM availability</i>					
Yelp star	-0.117 (0.076)	-0.224** (0.089)	0.043 (0.142)	-0.181** (0.077)	0.118 (0.180)
Yelp rating	-0.067 (0.350)	0.227 (0.409)	-0.986 (0.651)	0.141 (0.354)	-0.149 (0.866)
Yelp rating × Yelp star	0.490 (0.512)	0.293 (0.630)	1.709** (0.854)	0.438 (0.530)	-0.146 (1.136)
Observations	13,758	8,641	4,271	11,895	1,863
Sample	Full	100–500 reviews	500 + reviews	Not Michelin	Michelin
Mean 6 PM availability	0.745	0.797	0.634	0.794	0.416
Within-restaurant SD in availability	0.241	0.211	0.318	0.220	0.376
<i>Panel (b): 7:00 PM availability</i>					
Yelp star	-0.191** (0.092)	-0.339*** (0.102)	-0.005 (0.145)	-0.272*** (0.094)	0.095 (0.106)
Yelp rating	-0.022 (0.443)	0.690 (0.472)	-1.528** (0.743)	0.265 (0.442)	-0.283 (0.640)
Yelp rating × Yelp star	0.526 (0.658)	-0.180 (0.753)	2.483** (1.039)	0.525 (0.667)	-0.569 (0.733)
Observations	13,758	8,641	4,271	11,895	1,863
Sample	Full	100–500 reviews	500 + reviews	Not Michelin	Michelin
Mean 7 PM availability	0.586	0.664	0.412	0.656	0.114
Within-restaurant SD in availability	0.219	0.212	0.255	0.223	0.191
<i>Panel (c): 8:00 PM availability</i>					
Yelp star	-0.145* (0.084)	-0.210** (0.101)	-0.059 (0.156)	-0.237*** (0.082)	0.205 (0.138)
Yelp rating	-0.108 (0.389)	-0.034 (0.457)	-0.761 (0.753)	0.226 (0.359)	-0.662 (0.765)
Yelp rating × Yelp star	0.794 (0.590)	0.790 (0.686)	1.704 (1.061)	0.766 (0.557)	-0.109 (1.049)
Observations	13,758	8,641	4,271	11,895	1,863
Sample	Full	100–500 reviews	500 + reviews	Not Michelin	Michelin
Mean 8 PM availability	0.682	0.756	0.521	0.754	0.202
Within-restaurant SD in availability	0.226	0.205	0.296	0.222	0.257

별점 3.5, 4, 4.5 통합 결과

음식 유형별로 구분 분석

- 1) 전체
- 2) 리뷰 적음 (100~500)
- 3) 리뷰 많음 (500+)
- 4) 미셰린 아님
- 5) 미셰린

미셰린 식당은 컷오프
근처에서 별점 표시로 인한
불연속이 발생하지 않음



Conducting RDD Analysis

RDD분석을 어디에 활용해볼 수 있을지 생각해보자

■ 디지털 배지/인증 부여의 효과 분석

- 여러분의 회사가 e커머스 플랫폼 사업을 하고 있는데, 고객평이 좋고 판매량이 많은 판매자들에게 “**우수판매자**” 인증 배지를 부여한다고 해보자
✓ 해당 마크 부여가 판매 성과에 영향을 미칠까? 얼마나?

등급	판매조건		서비스만족도		
	판매건수	판매금액	배송처리율	품질취소율	상품만족도
최우수셀러	30건 이상	300만원 이상	의무상품발송기한 이내 처리 90% 이상	품질 비율 3% 미만	구매후기 평점 비율 90% 이상

할당변수 바뀌가며 분석...?

- ‘지역화폐 사용 제한(연 매출 30억원 이상 가맹점)’이 가맹점 매출에 미치는 영향, 영업 실적 기반 ‘올 해의 사원’ 선정이 해당 직원의 다음 해 실적에 미치는 영향, 구매 실적 기반 VIP 고객 선정 여부가 해당 고객의 다음 달 제품 구매량에 미치는 영향, 만 65세부터 지급되는 연금 지급이 노년층의 소비지출에 미치는 영향 등

Conducting RDD Analysis

보너스2) 교육복지 학교 지정이 학력격차 완화에 미치는 영향

$$Y_i(\text{학습부진아 비율}) = A_0 + A_1 \cdot T + A_2 \cdot (X-C)_i + A_3 \cdot (X-C)_i \cdot T + \varepsilon_i$$

- 저소득 학생 수(할당변수)가 40명 이상인 경우(컷오프) 교육복지학교에 선정됨 (결과변수: 학습부진아 비율)

<표 5> 구간 범위별 교복특 학교지정이 학습부진아 비율에 미치는 영향

과 목	변 수	단절점에서 구간(bandwidth) 범위				
		±10	±15	±20	±25	±30
영 어	교복특 학교 [T]	.30 (.38)	.34 (.38)	.46 (.38)	.45 (.35)	.52 (.35)
	저소득학생수 [X-40]	.03 (.04)	.07 (.02)	.06 (.01)	.05 (.01)	.04 (.01)
	교복특 학교*저소득학생수 [T·(X-40)]	-.21 (.08)	-.08 (.03)	-.08 (.03)	-.05 (.02)	-.04 (.02)
	절편 [A ₀]	2.47 (.19)	2.47 (.17)	2.41 (.16)	2.38 (.15)	2.31 (.14)
	n	103	139	199	235	269
수 학	교복특 학교 [T]	1.08* (.51)	.90 (.49)	1.27** (.42)	1.26** (.39)	1.27** (.40)
	저소득학생수 [X-40]	.02 (.04)	.05 (.03)	.07 (.01)	.06 (.01)	.05 (.01)
	교복특 학교*저소득학생수 [T·(X-40)]	-.06 (.11)	-.05 (.07)	-.10* (.03)	-.08* (.02)	-.08* (.02)
	절편 [A ₀]	3.81** (.28)	4.01** (.24)	3.77** (.19)	3.79** (.17)	3.72** (.17)
	n	103	139	199	235	270

주. ** $p < .01$ * $p < .05$

서로 다른
대역폭에 대해
결과를 나열

수학 교과목에서는
LATE가 통계적으로
유의하게 나타남

해당 연구에서는
교차항에도 관심이
있음 (왜?)

Comparing RDD with Other Methods

Conducting RDD Analysis

RDD vs PSM (1)

- PSM에서는 처치 여부 이외에 특성이 유사한 처치군과 대조군을 엄밀하게 구성하는 것이 목표
 - 혼동변수를 활용한 성향점수 추정 → 성향점수 기반 매칭 → 매칭된 처치군과 대조군의 유사도 검증(SMD, t-검정 등) → ...
 - 적절한 성향점수 추정모형 및 매칭 방식의 결정과 매칭 결과의 검증이 중요함
- RDD는 할당변수의 임계값을 기준으로 처치군과 대조군을 설정
 - 임계값 근처의 관측치들이 처치 여부를 제외하고는 유사한 특성을 가질 것이라고 가정
 - 적절한 할당변수, 임계값 및 대역폭 설정 및 주요 가정의 만족 여부를 따져보는 것이 중요함

Conducting RDD Analysis

RDD vs PSM (2)

- 연구 목적, 가용한 데이터 등에 따라 더 적절한 방법을 선택할 수 있음
 - 예를 들어, 처치군과 대조군에 대해 다양한 공변량의 확보가 어려운 경우 PSM의 활용이 어려울 수 있음
 - 반대로 컷오프를 명확하게 정의하기 어려운 경우 RDD의 활용이 어려울 수 있음
- RDD와 PSM 역시 상호보완적으로 활용될 수 있는 가능성이 있음
 - 예) RDD를 활용하려 했더니 임계값 근처의 관측치가 적어서 적절한 분석이 어려운 경우, 대역폭을 넓히는 방안을 고려해볼 수 있음
 - 그런데 대역폭이 넓어지면 처치군과 대조군의 동질성이 떨어짐
 - 대역폭을 넓히되 1차적으로 PSM을 통해 매칭된 정제된 표본에 대해 RDD 적용?

Conducting RDD Analysis

RDD vs DID

- RDD와 DID 역시 서로의 주요 관심사가 다르고 각자의 장단점이 존재하는 방법론이지만, **할당변수가 시간인 경우** 두 방법이 **병행 활용** 가능하거나 심지어 **결합 활용**이 가능할 수도 있음
- 예를 들어, 이전에 살펴보았던 Card-Krueger의 최저임금 인상 사례에서, 뉴저지의 최저임금은 1992년 4월을 기준으로 이루어졌음
 - 만약 대상 지역에 대해 **월(주)별로** 고용 자료가 존재한다면, RDD를 활용하여 1992년 4월을 기점으로 단절이 발생했는지 분석해볼 수 있음
- 두 모형을 결합하여 활용하는 사례도 존재함
 - 코로나 바이러스 발발 시 주요 항만 락다운에 의한 효과를 분석함에 있어 RDD를 이용하여 단기(국지적) 효과를, DID를 이용하여 장기효과 분석
 - Bai, X., Xu, M., Han, T., & Yang, D. (2022). Quantifying the impact of pandemic lockdown policies on global port calls. *Transportation Research Part A: Policy and Practice*, 164, 224-241.

Recap

Recap

RDD의 주요 개념

- 임계값(cutoff, cutoff)
 - 처치를 받게 되는 조건을 결정하는 기준점
 - ✓ 예) Yelp 평점에서 3.25, 3.75, 4.25, 4.75 등, 장학금 지원 기준 점수
 - 임계값을 기준으로 처리군과 대조군이 구분됨
- 할당 변수(assignment variable, running variable)
 - 임계값 설정의 대상이 되는 변수
 - ✓ 예) Yelp 평점, PSAT 점수 등
 - cf) 결과 변수(outcome variable): 처치에 따라 영향을 받을 것으로 기대하는 변수
 - ✓ 예) 식당 예약률, 대학 졸업률 등
- 대역폭(bandwidth)
 - 배정변수의 특정 임계값 기준으로 어느 범위의 데이터를 포함할 것인지를 나타내는 기준
 - ✓ 예) Yelp 평점 ± 0.25 점, PSAT 점수 ± 1 점 등

Recap

LATE

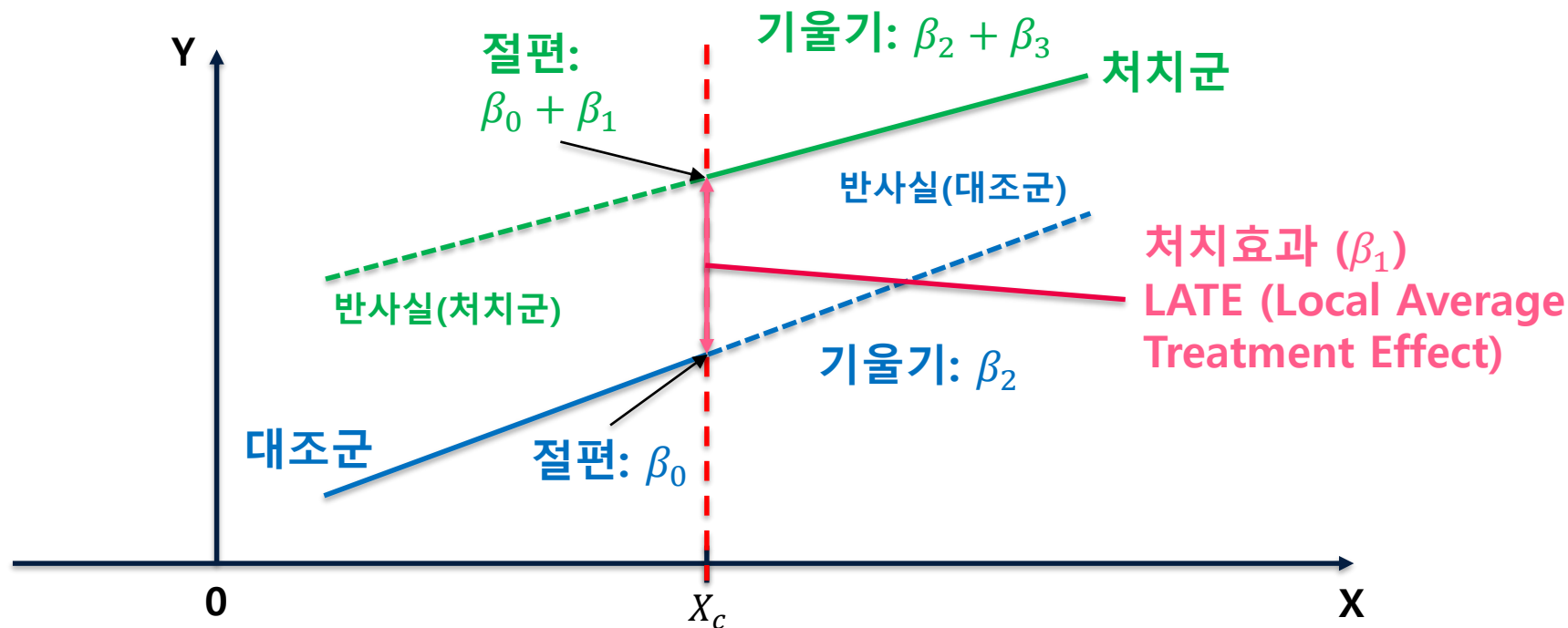
- RDD에서는 컷오프를 기준으로 바로 위와 바로 아래에 있는 관측치의 비교를 통해 처치 효과를 추정함
 - 이를 위해 처치군과 대조군의 추세를 선형회귀모형을 이용해 분석하고 컷오프 근처에서의 차이를 비교함
- 이러한 효과는 임계값 근처의 관측치에 집중하므로, 이를 **지역 평균 처리 효과(LATE, Local Average Treatment Effect)**라고 부르기도 함
 - 임계값 근처에서만 국지적으로 유효한 처치 효과임
 - 컷오프 값이 변경되었을 때에도 해당 처치효과가 유지된다고 보기 어려울 수 있음

Recap

RDD 추정 결과 시각화

- 아래 식 추정 결과에 따라....

$$Y_i = \beta_0 + \beta_1 \text{Treat}_i + \beta_2 X_i' + \beta_3 \text{Treat}_i X_i'$$



Recap

RDD 분석의 주요 단계

- RDD 분석의 주요 단계는 다음과 같이 설정할 수 있음
 1. 할당변수 및 임계값 결정
 2. 대역폭 및 모형 선택
 3. RDD 추정 및 결과 해석

Recap

RDD vs PSM vs DID

- PSM에서는 처치 여부 이외에 특성이 유사한 처치군과 대조군을 엄밀하게 구성하는 것이 목표
 - 혼동변수를 활용한 성향점수 추정 → 성향점수 기반 매칭 → 매칭된 처치군과 대조군의 유사도 검증(SMD, t-검정 등) → ...
 - 적절한 성향점수 추정모형 및 매칭 방식의 결정과 매칭 결과의 검증이 중요함
- RDD는 할당변수의 임계값을 기준으로 처치군과 대조군을 설정
 - 임계값 근처의 관측치들이 처치 여부를 제외하고는 유사한 특성을 가질 것이라고 가정
 - 적절한 할당변수, 임계값 및 대역폭 설정 및 주요 가정의 만족 여부를 따져보는 것이 중요함
- RDD와 DID은 각자의 주요 관심사가 다르고 각자의 장단점이 존재하는 방법론이지만, 할당변수가 시간인 경우 두 방법이 병행 활용 가능하거나 심지어 결합 활용이 가능할 수도 있음