

# 인과추론과 실무 : RDD

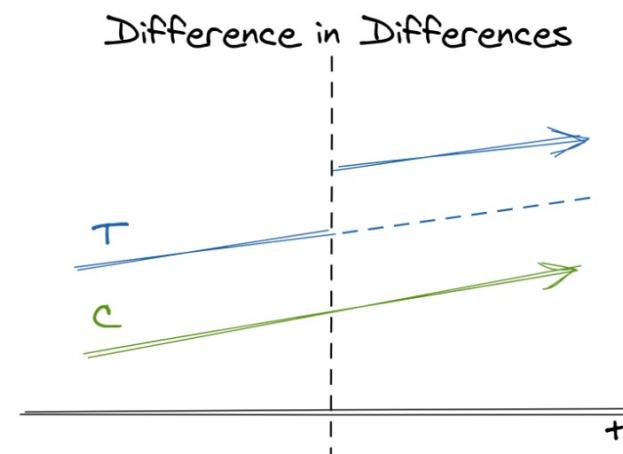
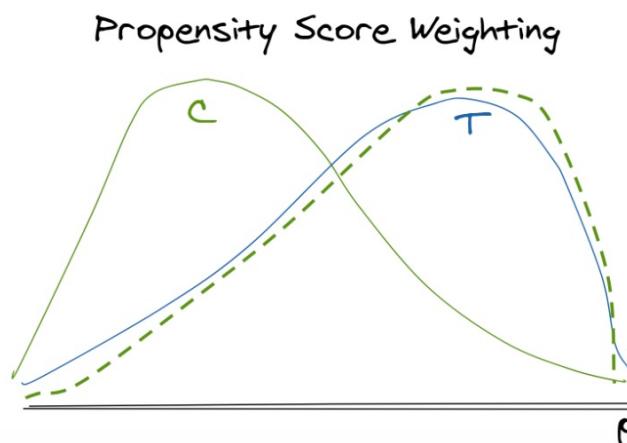
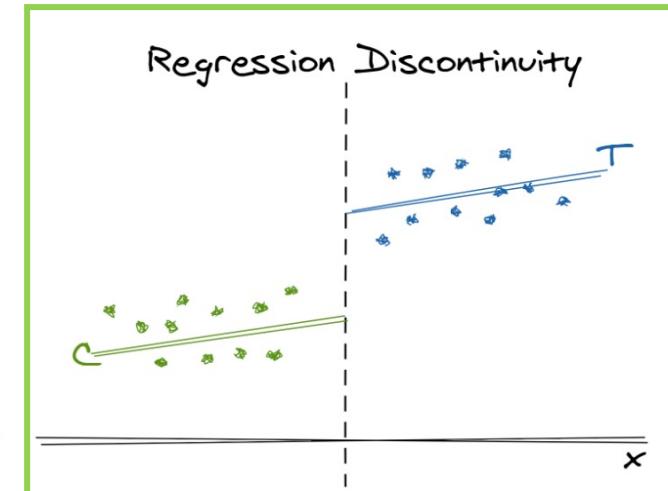
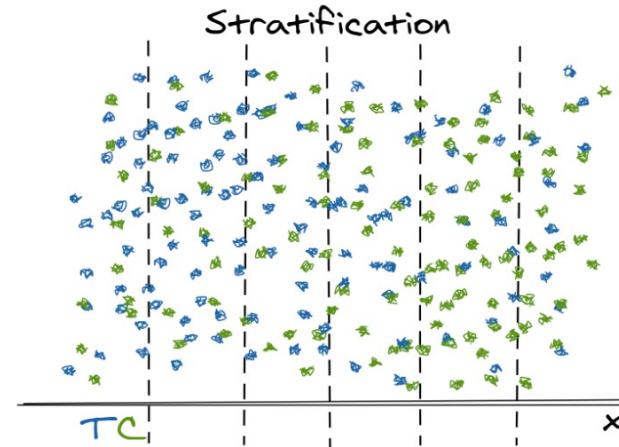
가짜연구소 인과추론팀

발표자 : 손영주

# 1. RDD(회귀 불연속 설계)

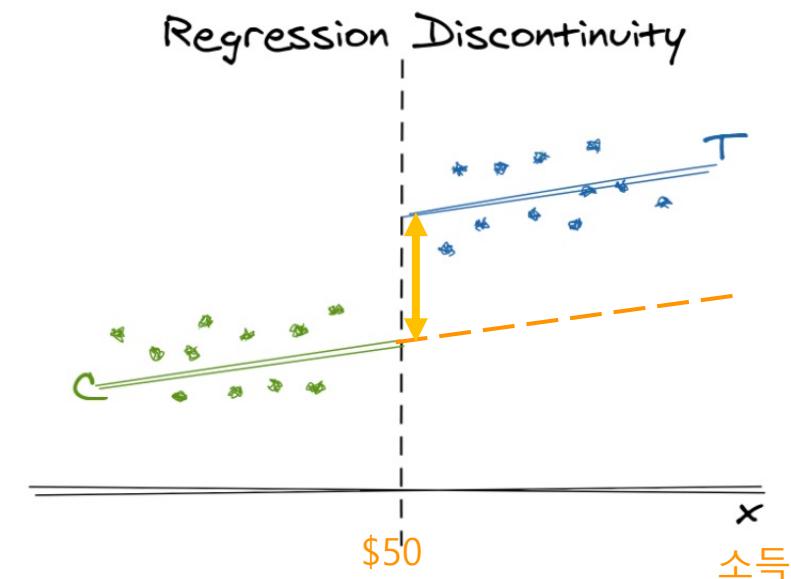
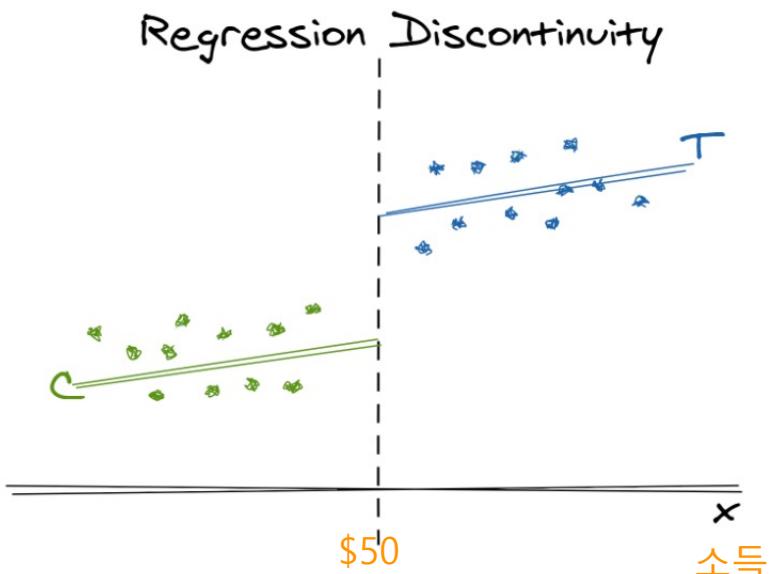
# Regression Discontinuity Design

- 처치 배정에 인위적인 불연속성(discontinuity)을 활용하여 처치효과를 식별



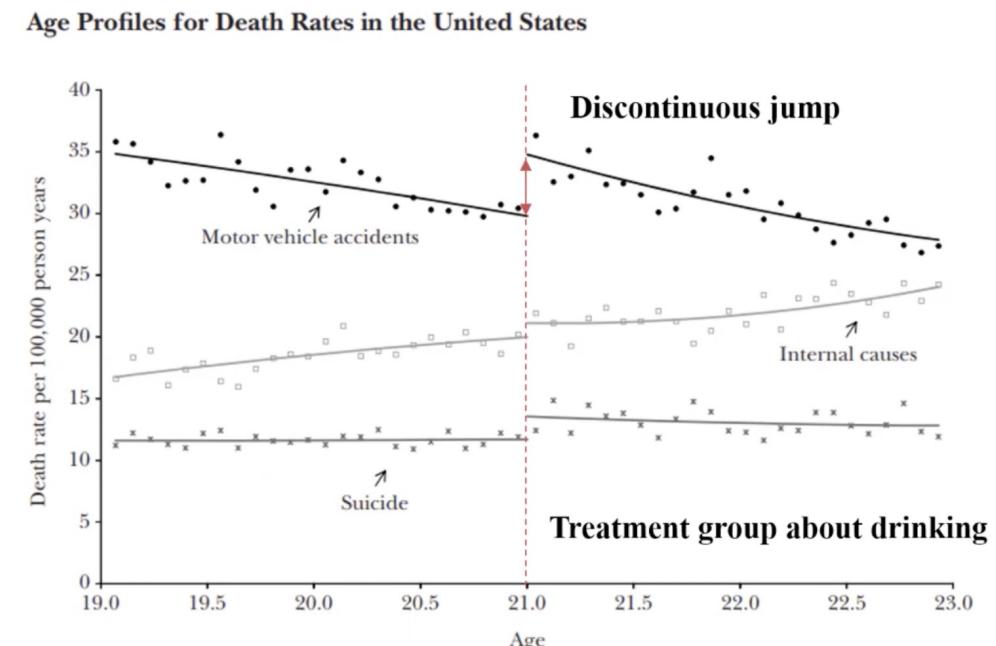
# Regression Discontinuity Design

- 특정 상황에서 불연속적으로 일어난 변화를 포착하여, 마치 실험에서 무작위 배정한 것처럼 활용
- 유사한 두 그룹이 임곗값을 기준으로 나뉠 때 사용할 수 있는 Quasi-Experiment 방법
  - 예 1. 소득 기준의 정부 보조금 지원 프로그램: 일소득 \$50 미만 가정만 매월 \$200 지급하는 경우 프로그램 효과
  - 예 2. 은행 고객의 계좌 잔고가 \$5,000 이하인 경우 신용카드 발급에 수수료 부과하는 경우 카드 발급 효과
- 소득 \$49.9와 \$51.1, 잔고 \$4,999와 \$5,001은 유사한 그룹이었으나 우연히 cutoff에 의해 다른 그룹으로 분류되어 서로 다른 처치를 받음(또는 처치 받을 가능성에 차이 발생)



# Regression Discontinuity Design

- Running Variable(Assignment Variable, Forcing Variable) 과 Cutoff
  - Running Variable = Cutoff를 결정하는 변수
  - Running Variable 자체는 결과변수(outcome)와 연관이 있을 수 있지만(직접 연관되거나, confounder 통해)
  - cutoff 자체는 결과와 관련없이 임의로 정해진 값
- 예: 음주와 건강에 관한 연구에서 cutoff = 법정 음주 가능 연령(미 21세), running variable = 연령. 연령과 건강은 연관이 있을 수 있으나 음주 가능 연령 자체는 arbitrary value



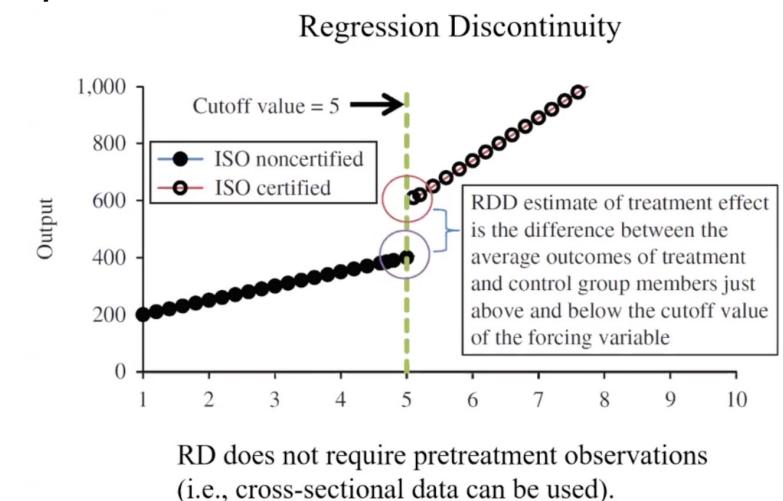
# RDD와 DiD(Difference in Difference, 이중차분법)

데이터가 지닌 맥락에 맞는 가정을 가진 방법론 선택:

둘 다 사용 가능한 경우 모두 확인하면 가장 좋고, DiD의 가정을 검증하는 것이 비교적 쉬움

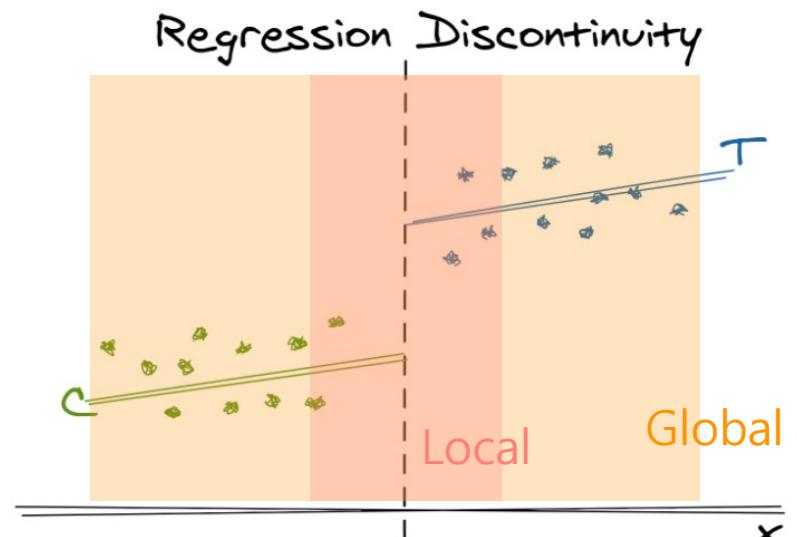
- (1) ISO 인증이 기업의 성과에 미치는 효과 (Ho et al.2017)
  - A. ISO 인증 여부만 알고 있을 경우: 인증 여부에 영향을 미치는 변수를 covariate으로 통제
  - B. ISO 인증 통과 여부를 결정하는 연속적인 점수와 그 cutline을 알고 있는 경우: RDD
  - C. ISO 인증 이전과 이후의 성과에 대해 패널 데이터를 갖고 있는 경우: RDD, DiD

- (2) 셧다운제도 적용이 게임 플레이시간에 미치는 효과 (Jo et al.2020)
  - 16세라는 나이를 기준으로 cutoff되어 RD로도 분석할 수 있으나 전후 패널 데이터를 활용하여 DiD분석 진행



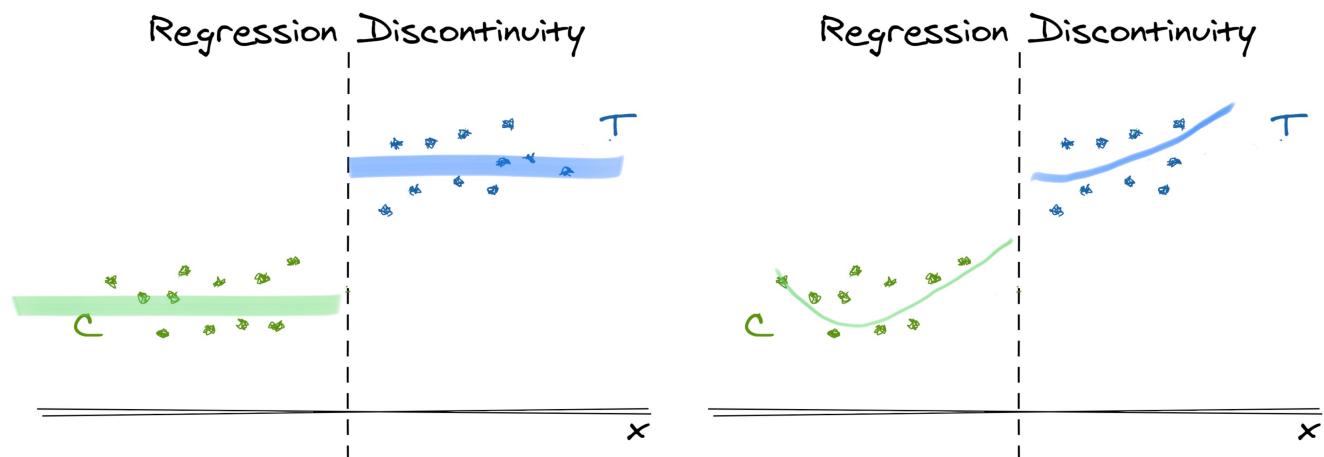
# RDD와 Bandwidth

- Cutoff '주변'을 어느 범위까지 고려할까?
  - (1) Global Strategy: Running Variable의 모든 구간을 활용하여 분석
  - (2) Local Strategy: Running Variable의 cutoff 주변의 일정 구간을 잘라서 분석
- Tradeoff
  - Local → 샘플사이즈 감소, variance 증가 → efficiency 감소
  - Global → selection bias 가능성 증가
- MSE(Mean Squared Error)
  - variance와 bias의 합을 최소화하는 bandwidth 결정



# RDD와 RV 모델링

- 두 집단을 비교하는 기준
  - (1) Non-Parametric: 동질적인 두 집단이 cutline을 기준으로 나뉘었다고 보고(control vs treatment) 단순히 cutoff 전후의 평균을 비교하여 두 그룹의 차이 확인
  - (2) Parametric: Linear, quadratic 등 control function을 가정하여 running variable과 outcome의 관계를 통제하고자 함



# RDD 설계 Choices: Bandwidth x RV 모델링

- Modeling choices in RD

		Bandwidth	
		Local	Global
Modeling of Running Variable (Control Function)	Nonparametric	Local Nonparametric (Local Experiment)	Global Nonparametric (Global Experiment)
	Parametric	Local Parametric (Local Regression)	Global Parametric (Global Regression)

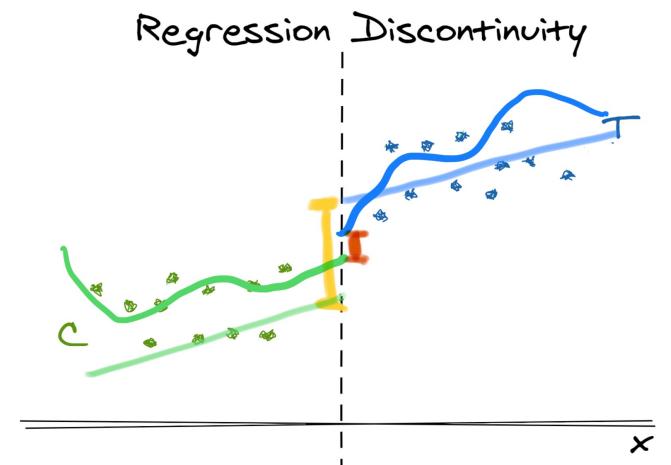
- 여러 전략을 섞어서 사용하는 것이 기본(\*Global Nonparametric)
  - 유의성 확보를 위해 Global Bandwith로 샘플 수를 늘리면 bias 가능성이 증가하기 때문에, running variable의 parametric function을 설정해서 bias를 최대한 보정하고자 하게 됨

## RDD의 가정들

- No Sorting / No-Manipulation-With-Precision
  - Cutline 배정을 대상자가 스스로 결정하거나, 일부러 이동하는 현상(self-selection)이 없음
    - cutoff 규칙이 대상자에게 알려지지 않았거나, 알았더라도 이동할 수 없음
    - 예: 혜택(보조금)을 제공하는 곳으로 이동
  - Test: Density Discontinuity
    - 임곗값(cutoff) 근처에서 실험 대상의 밀도가 증가한다면 self selection이 일어난 것이므로 RDD 가정 위배
- Bunching Analysis
  - 대상이 가로축(running variable) 변수를 조작할 수 있는 상황에서, cutoff에 반응하여 변동이 일어났는지 분석

# RDD의 가정들

- (Parametric Strategy 사용 시 가정): Untreated – Treated 그룹 간의 variance는 running variable에 대한 function으로 모두 설명할 수 있고, 해당 함수를 통해 차이를 통제하면 두 그룹이 서로 비교 가능해짐
  - Control function의 형태에 따라 따라 결과(인과추정량)이 달라지는데, 실제 형태가 어떠한지 현실 세계에서는 알 수 없다
  - Bandwidth를 줄이면 function 형태에 따른 차이가 줄어들지만 sample size 감소 -> MDE
  - 여러 형태의 control function을 test – 최대 3차식



## RDD의 가정들

- **Covariate Balance**

- 두 집단의 Covariate들이 running variable의 임곗값(cutoff) 근처에서 연속적이다
- 역시 직접 검증하는 것은 불가능하나(관찰 불가능한 covariate 존재), 관찰 가능한 covariate에 대해서 확인하면서 cutoff로 인해 나뉜 두 집단이 동질적이라는 가정을 보강
- Covariate와 running variable이 무관할 필요는 없고 cutoff 근처에서 연속적임을 보여준다

## RDD의 가정들

- 기대효과를 나타내는 변수를 연속적인 함수로 모델링할 수 있음
- 결과변수와 의사결정변수 사이에 단순한 모델을 적합시킬 수 있음

# RDD Regression Model

- (1) Nonparametric
- (2) Parametric
  - (2-1) Linear & Same model around the cutoff
  - (2-2) Linear & Different models around the cutoff
  - (2-3) Quadratic & Same model around the cutoff
  - (2-4) Quadratic & Different model around the cutoff
  - ...

# RDD Regression Model

- (1) Nonparametric

$$Y = \beta_0 + \beta_1 \text{AboveCutoff} + \text{Controls} + \varepsilon \quad \leftarrow \text{treatment 평균 vs control 평균의 차이 } \beta_1$$

*AboveCutoff: 1 or 0*

- (2) Parametric

- (2-1) Linear & Same model around the cutoff

$$Y = \beta_0 + \beta_1 \text{AboveCutoff} + f(\text{rv}) + \text{Controls} + \varepsilon$$

- (2-3) Linear & Different models around the cutoff

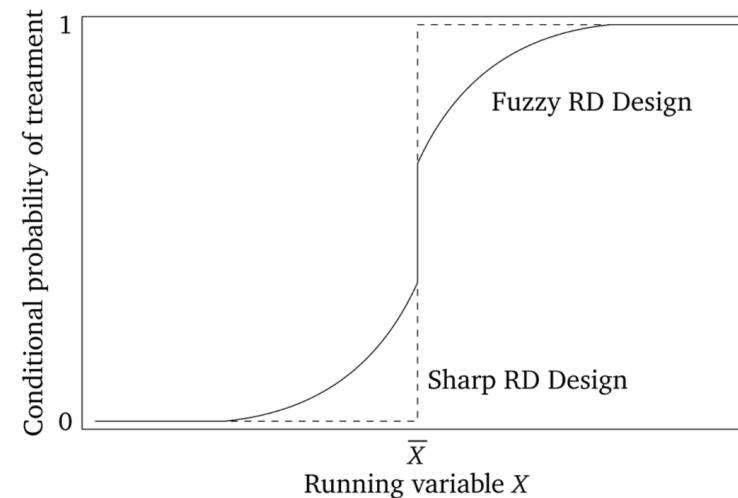
$$Y = \beta_0 + \beta_1 \text{AboveCutoff} + f(\text{rv}-\text{cutoff}) + f'(\text{rv}-\text{cutoff}) \times \text{AboveCutoff} + \text{Controls} + \varepsilon$$

*1*  $Y = \beta_0 + \beta_1 + f(\text{rv}-\text{cutoff}) + f'(\text{rv}-\text{cutoff}) + \text{Controls} + \varepsilon$

*0*  $Y = \beta_0 + f(\text{rv}-\text{cutoff}) + \text{Controls} + \varepsilon$

## RDD의 유형: Sharp vs Fuzzy

- (1) Sharp RDD: 임곗값을 기준으로 처치 여부(0,1)가 100% 결정됨
- (2) Fuzzy RDD: 임곗값을 기준으로 처치를 받을 확률이 변화하며, 미처치 개체가 존재할 수 있음



## Additional Robustness Check

- Placebo Tests
  - 임계값 주변에 존재하지만 처치를 받지 않은(< Fuzzy Design) 데이터 포인트를 대상으로 불연속성이 나타나는지 확인
    - 불연속성이 나타난다면, 관찰되지 않는 다른 confounding variables 또는 potential bias 존재할 가능성
    - 불연속성이 없고, treatment threshold 근처에서만 jump가 일어난다면 이 jump처치 때문이라는 증거를 보강

## RDD의 강점과 약점

- 강점
  - RCT의 비용과 노력 없이 이미 확보된 데이터로 분석할 수 있음(정책, 비즈니스 결정으로 인한 불연속: 지리적, 연령, 시간)
  - 시각적인 설명이 가능
  - treatment 이전 기간의 pre-treatment outcome을 요구하지 않음 (c.f. 패널 데이터 분석, DiD)
- 약점
  - 가정 입증의 어려움(c.f. DiD)
  - 외적타당성을 확보하기 위해서는 추가적인 가정이 필요 (일반적으로는 경계선 근처에 대한 인과관계 파악)

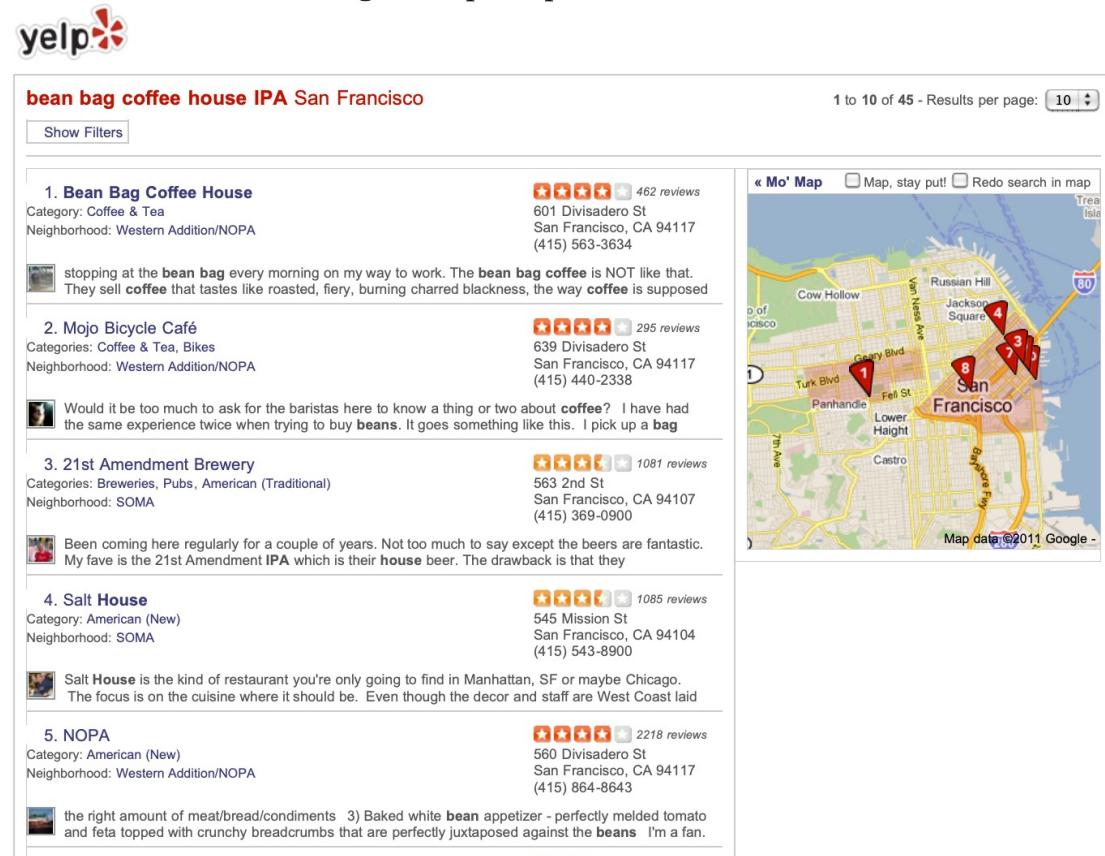
## 2. RDD with Industry Case

# (1) 별점이 식당 예약에 미치는 효과 (Yelp)

Q.Yelp.com에서의 별점이 높으면 식당의 인기가 많아질까?

- 데이터(1) Yelp review of San Francisco, California(2011 기준)
- 데이터(2) 온라인 식당 예약 사이트 availability 데이터, 2010.07~10
- Outcome: 예약 가능성
- Running Variable: 평균 Yelp 평점
- Cutoff: .5점 단위 (0.5, 1.0, 1.5, 2.0, .., 4.5)

Fig. 1: Sample Yelp.com Search Results



## (1) 별점이 식당 예약에 미치는 효과 (Yelp)

[Yelp.com에서의 별점이 높으면 식당의 인기가 많아질까?]

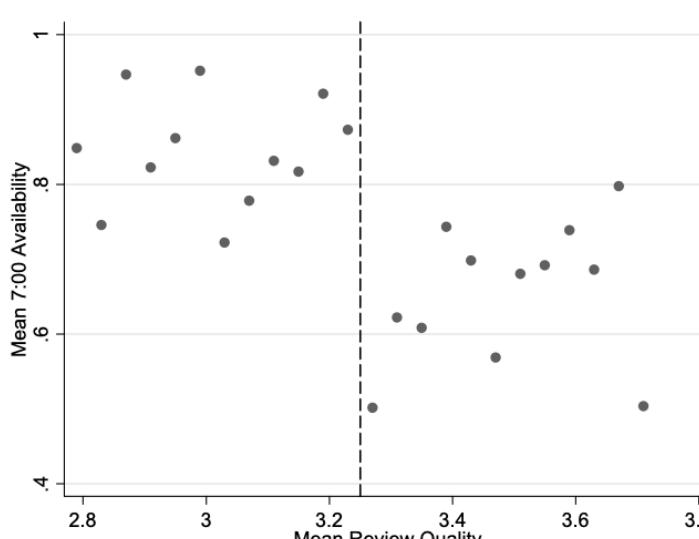
$$y_{it} = \alpha + \beta * DR_{it} + \gamma f(R_{it}) + \varepsilon_{it}$$

$$y_{it} = \alpha + \beta * I(R_{it} > \hat{R}) + \gamma_1 * \tilde{R}_{it} + \gamma_2 * \tilde{R}_{it} * I(R_{it} > \hat{R}) + \varepsilon_{it}$$

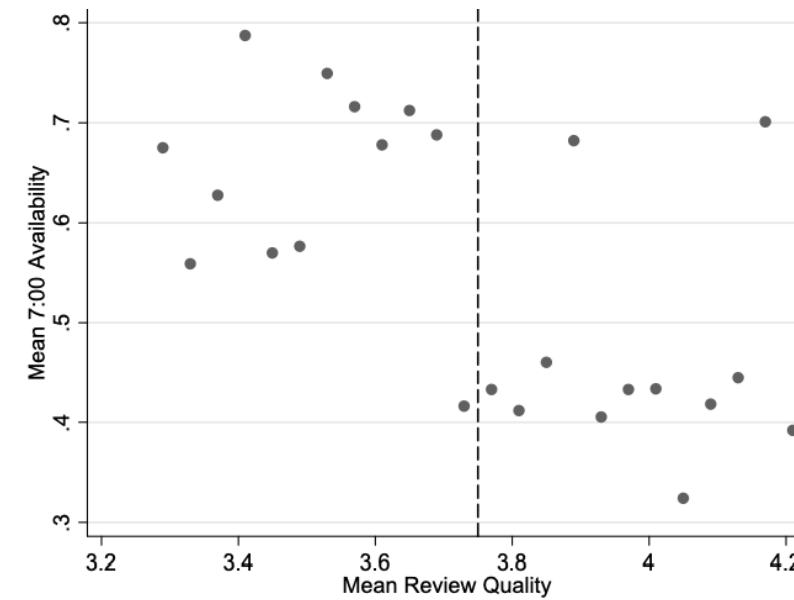
- $y_{it}$  : 식당 i에 t날짜에 4인 예약 가능한지 여부를 나타내는 indicator
- $DR_{it}$ : Yelp가 식당 옆에 노출하는 별점(0.5단위의 step function: sharp RD)
- $R_{it}$  : Yelp 실제 식당 평균 리뷰점수, 임곗값으로 정규화
- $\hat{R}$  : 0.5 star threshold
- $\tilde{R}_{it} = R_{it} - \hat{R}$
- 리뷰의 품질과 식당의 품질은  $f$ 로 모두 계산되며, RD 임곗값에서 불연속적으로 변하지 않는다고 가정
- Bandwidth: 노출 평점이 변하는 구간(0.5단위)

# (1) 별점이 식당 예약에 미치는 효과 (Yelp)

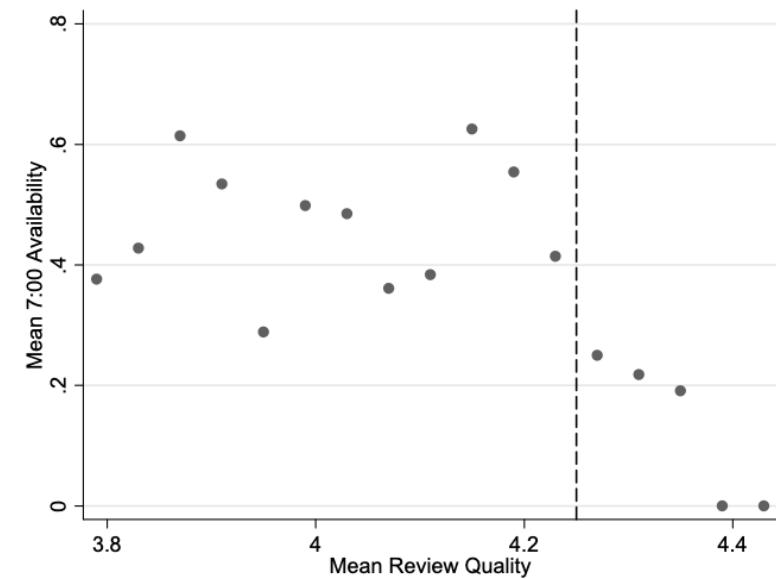
Fig. 2: Reservation Availability at 7:00 pm by Average Yelp Rating



90% -> 70%



70% -> 45%



45% -> 20%

# (1) 별점이 식당 예약에 미치는 효과 (Yelp)

Table 2: Regression Discontinuity Results at Individual Thresholds

Yelp Display Rating	6:00 Availability			7:00 Availability			8:00 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 1. Contains RD estimates of the effects of an additional Yelp half-star on availability

2. Availability measures indicate whether the reservations were available at that time on Thursday, Friday, or Saturday when queried 36 hours in advance

3. Standard errors are clustered at the restaurant level

4. Stars denote significance levels: 10% (\*), 5% (\*\*), and 1% (\*\*\*)

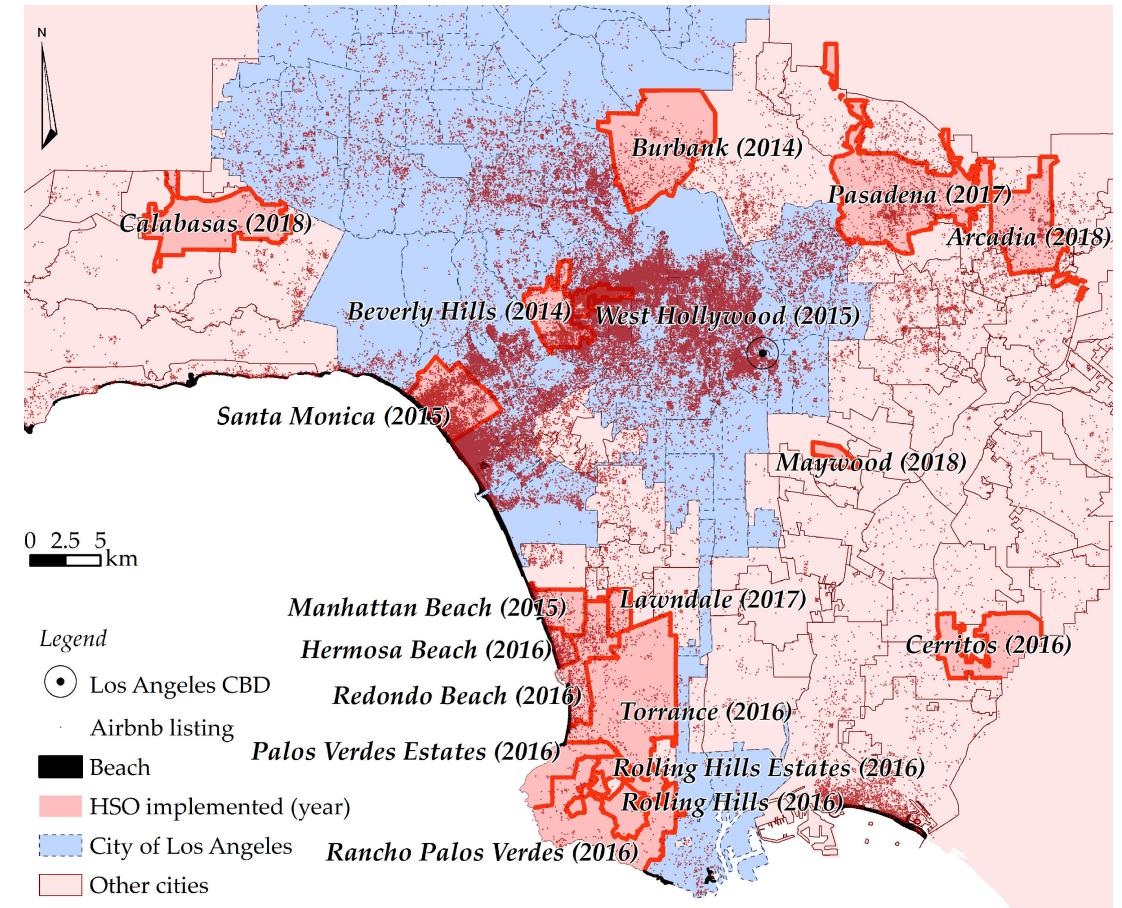
## (1) 별점이 식당 예약에 미치는 효과 (Yelp)

- Robust test
  - Yelp 리뷰 임계값에서 식당 밀도가 불연속적이지 않음
  - Yelp 리뷰 임계값에서 리뷰어 특성이 불연속적이지 않음
  - Yelp 리뷰 임계값을 넘는다고 해서 다른 예약 데이터베이스 서비스의 평가에서 확인한 음식의 질, 식당의 질에 변화가 나타나지 않음

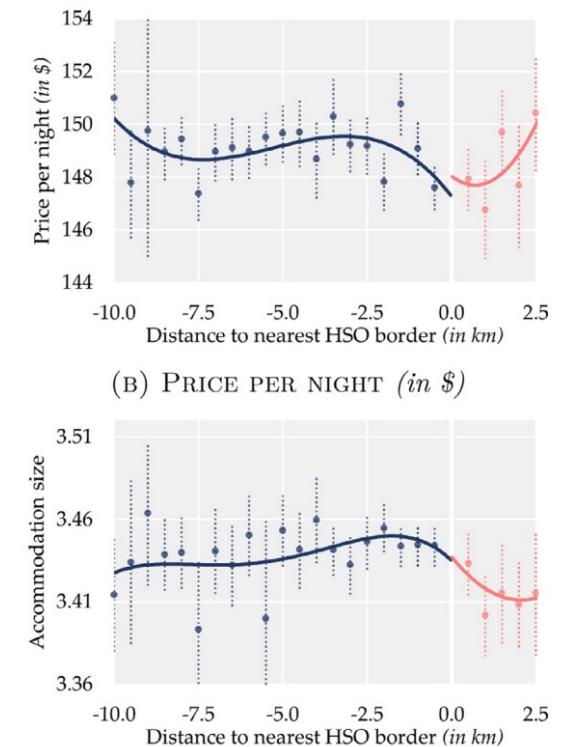
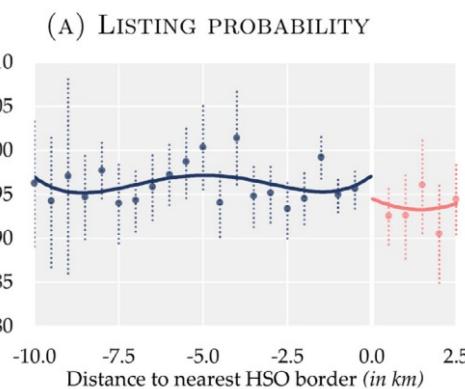
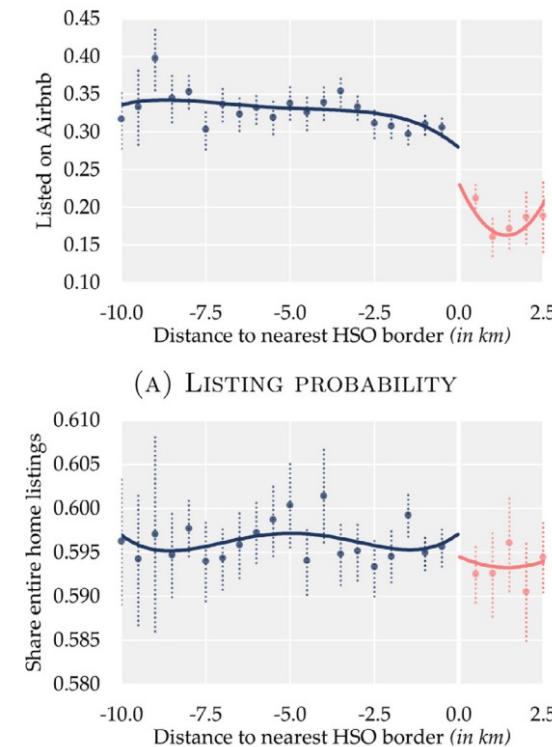
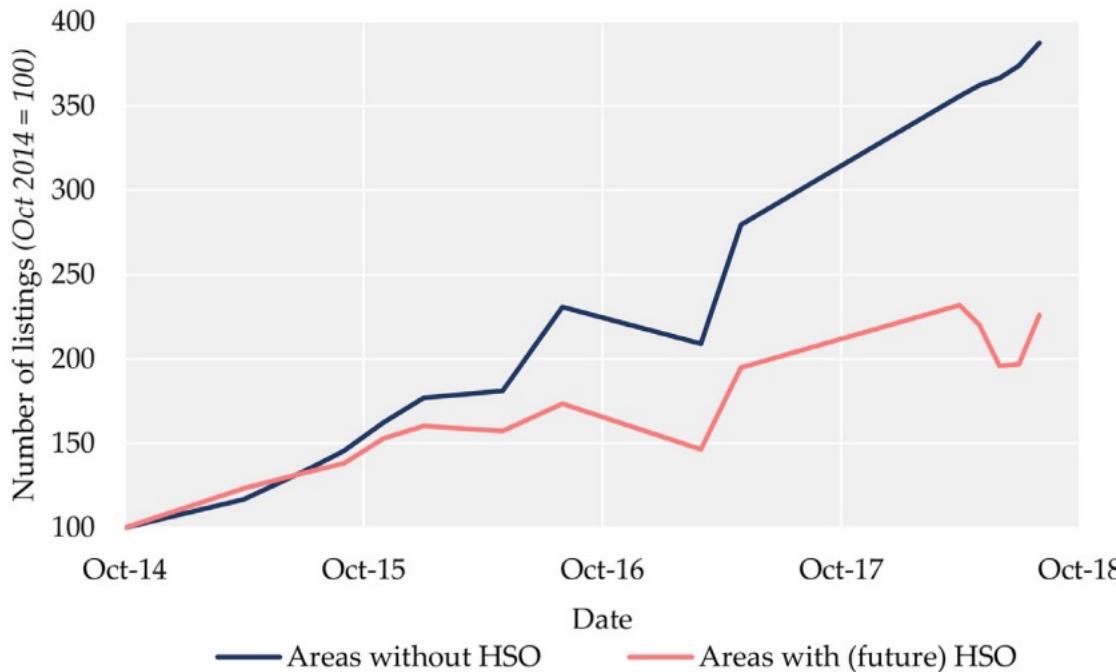
## (2) 단기임대 규제가 주거 시장에 미치는 효과 (Airbnb)

Q. Airbnb를 규제하면 주택가격이 낮아질까?

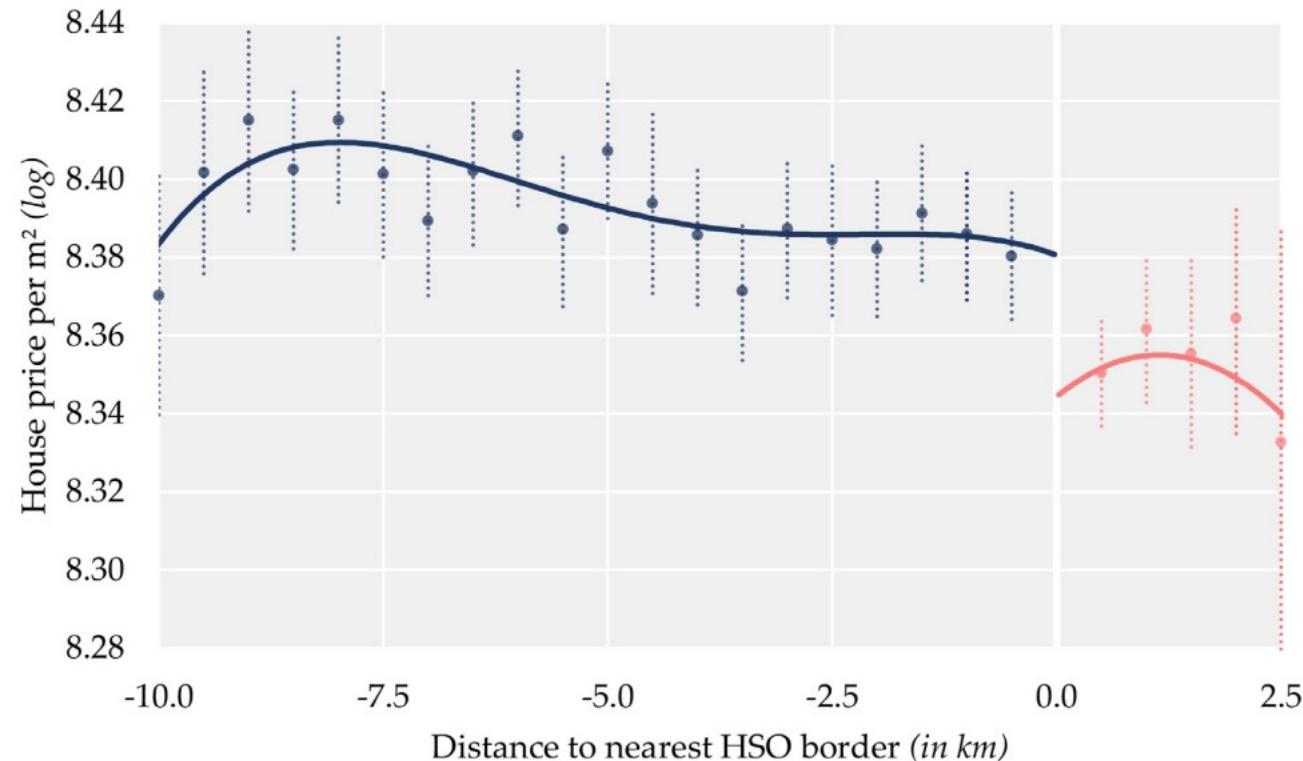
- HSO(Home-Sharing Ordinances): 관광지 주택가격 상승으로 규제 도입
  - 88개 중 18개 도시가 HSO를 도입(우측 그림의 붉은 색)하여 단기임대 세금 부과 및 금지(일부지역)
- 공간 회귀 불연속설계
- Outcome: Airbnb 목록, 주택가격, 임대료
- Running Variable: HSO 경계까지의 거리
- Cutoff: HSO지역( $RV=0$ )



## (2) 단기임대 규제가 주거 시장에 미치는 효과 (Airbnb)



## (2) 단기임대 규제가 주거 시장에 미치는 효과 (Airbnb)



## (2) 단기임대 규제가 주거 시장에 미치는 효과 (Airbnb)

**Table 4**  
Baseline results for Airbnb listings.

	<i>Panel RDD</i> (1)	+ Border segment f.e. (2)	Bandwidth: $h^* \times 2$ (3)	Bandwidth: $h^*/2$ (4)	Bandwidth: $h^*/5$ (5)	Rooms not allowed (6)	Measurement error (7)
<i>Panel A: Entire homes/apartments</i>							
HSO implemented	-0.0614*** (0.0122)	-0.0696*** (0.0123)	-0.0815*** (0.0110)	-0.0444*** (0.0154)	-0.0290 (0.0226)		-0.1058*** (0.0168)
HSO implemented $\times$ rooms allowed						-0.0682*** (0.0205)	
HSO implemented $\times$ rooms not allowed						-0.0701*** (0.0127)	
Spatio-temporal trend variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Property fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
HSO area $\times$ month fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Border segment $\times$ month fixed effects	No	Yes	Yes	Yes	Yes	Yes	Yes
Number of observations	270,906	270,621	425,117	154,015	80,896	270,741	253,448
Bandwidth, $b$ (in km)	1.6716	1.6708	3.3416	0.8354	0.3342	1.6712	1.9639
$R^2$	0.3481	0.3515	0.3550	0.3481	0.3439	0.3514	0.3546
<i>Panel B: Rooms</i>							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
HSO implemented	-0.0253 (0.0170)	-0.0363** (0.0172)	-0.0436*** (0.0156)	-0.0296 (0.0226)	-0.0082 (0.0315)		-0.0551** (0.0224)
HSO implemented $\times$ rooms allowed						0.0309 (0.0264)	
HSO implemented $\times$ rooms not allowed						-0.0595*** (0.0187)	
Spatio-temporal trend variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Property fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
HSO area $\times$ month fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Border segment $\times$ month fixed effects	No	Yes	Yes	Yes	Yes	Yes	Yes
Number of observations	171,778	171,448	259,880	94,365	45,267	171,433	156,710
Bandwidth, $b$ (in km)	1.815	1.812	3.3384	0.8346	0.3338	1.8117	2.0061
$R^2$	0.3339	0.3438	0.3424	0.3524	0.3558	0.3439	0.3482

Notes: We exclude within 200m of the borders of HSO areas in column (7). Standard errors are clustered at the census block level and in parentheses. \*\*\*  $p < .01$ , \*\*  $p < .05$ , \*  $p < .10$ .

**책**

- 실무로 통하는 인과추론 with 파이썬

**강의**

- YouTube [인과추론의 데이터과학] 채널 / 10-1~4강 회귀불연속설계

**블로그**

- <https://henamsingla.medium.com/regression-discontinuity-design-rdd-an-emerging-tool-for-causal-inference-968c16d47ce6>
- <https://playinpap.github.io/causal-inference-regression-discontinuity/>
- <https://assaeunji.github.io/statistics/2021-02-21-causal2/>

**논문**

- 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.
- Koster, H. R., Van Ommeren, J., & Volkhausen, N. (2021). Short-term rentals and the housing market: Quasi-experimental evidence from Airbnb in Los Angeles. *Journal of Urban Economics*, 124, 103356.
  - <https://www.sciencedirect.com/science/article/pii/S0094119021000383>

# 감사합니다

## Q&A