Pseudo Lab Causal Inference Lab

인과추론과 실무 : 불응과 도구변수

가짜연구소 인과추론팀

발표자 : 유도영

1. 불응과 도구변수

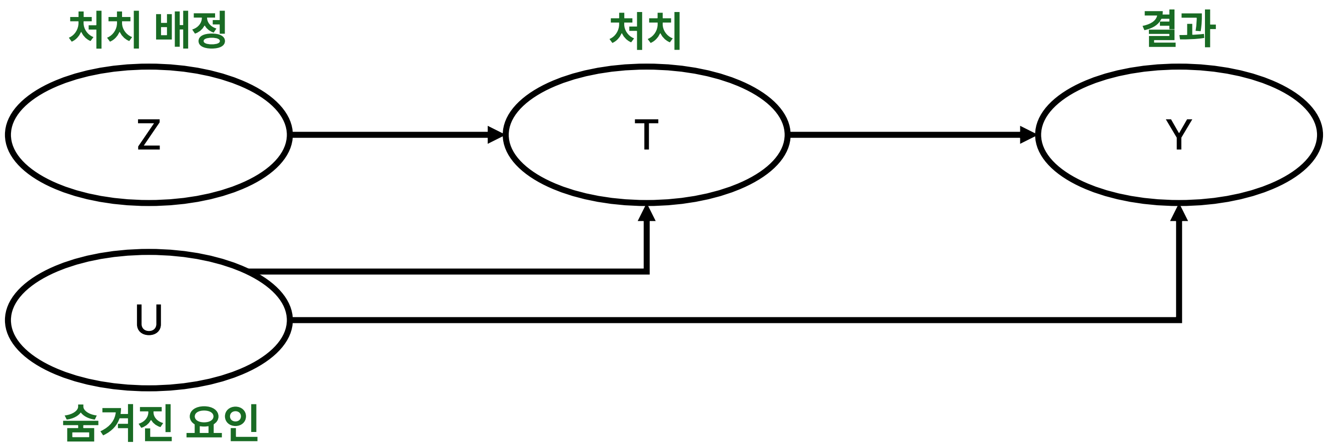
불응

- 회사가 서비스나 제품의 제공 여부를 **무작위**로 결정하더라도, 고객에게 이를 강요할 수 없음
- 즉, 처치를 배정받은 모든 사람이 처치 받지 않을 수 있으며, 이를 **불응** (non-compliance)이라고 함
- 처치 **배정**과 처치 **적용**을 분리하면 다음과 같이 네 그룹으로 분리할 수 있음
    - **순응자** (complier): 자신에게 배정된 처치를 받는 사람
    - **항시 참여자** (always-taker): 배정과 관계 없이 항상 처치 받는 사람
    - **항시 불참자** (never-taker): 배정과 관계 없이 처치를 한 번도 받지 않은 사람
    - **반항자** (defier): 배정된 처치와 반대되는 처치를 받는 사람

		처치 배정	
		No	Yes
처치 적용	No	순응자 & 항시 불참자	반항자 & 항시 불참자
	Yes	반항자 & 항시 참여자	순응자 & 항시 참여자

불응

- 불응을 DAG 형태로 나타낼 수도 있음



- 이 때 Z는 **도구 변수** (IV: Instrumental Variable)로서,
  1. 교란 없이 처치에 영향을 주고
  2. 처치를 거치지 않으면 결과에 영향을 미치지 않음

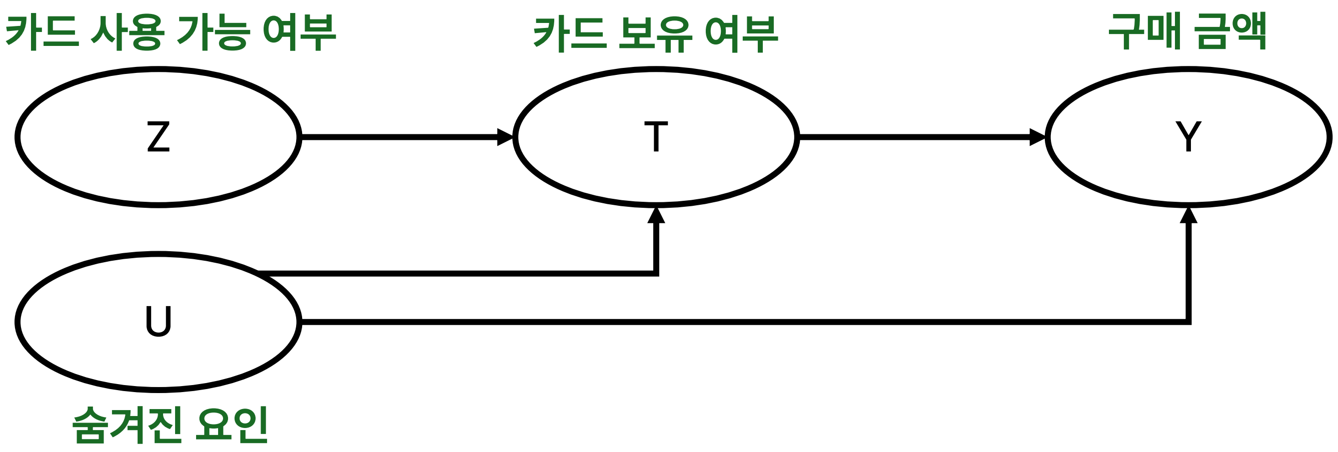
- 추가적인 가정이 없다면, U를 통과하는 뒷문 경로 때문에 처치가 결과에 미치는 영향을 식별할 수 없음
  - 이 효과를 식별하려면, 도구 변수를 현명하게 사용해야 함!

프라임 신용카드 예시

- 은행이 고객에게 프라임 신용카드를 제공했을 때, 어떤 영향을 미치는지 알고 싶어한다고 가정하자
  - 프라임 서비스를 제공하려면 큰 비용이 들지만 프라임 고객의 구매 금액이 500달러 이상 증가한다면 가치 있음
  - 은행은, 프라임 카드가 고객의 구매금액을 얼마나 늘리고 싶어하는지 알고 싶어함

- 은행은 10,000명의 고객을 대상으로 프라임 신용카드 이용 가능 여부를 무작위로 배정
  1. 절반은 실험군, 나머지 절반은 대조군으로 배정
  2. 하지만, 은행이 고객에게 카드를 선택하도록 강요할 수 없으므로 불응이 존재함

- 예시를 DAG 형태로 나타내면 다음과 같음



잠재적 결과 확장

- Z는 T의 원인이므로, **잠재적 처치** (potential treatment) T_Z를 정의할 수 있음
  - 그리고 잠재적 결과는 도구 변수 Z 에 대한 새로운 반사실 $Y_{z,t}$를 가짐

- 앞선 예시에서 Z는 무작위 배정되므로, Z가 Y에 미치는 영향인 처치 의도 효과 (ITTE: intent-to-treat effect)는 쉽게 식별 가능
  - $ITTE = E[Y|Z = 1] - E[Y|Z = 0] = E[Y_{1,t} - Y_{0,t}]$
  - 또한, 처치 의도 효과는 단순선형회귀 (구매 금액 ~ 카드 사용 가능 여부)로 추정할 수 있음

- 은행의 주요 목표는 프라임 카드의 혜택이 비용을 상회하는지 결정하는 것 ➡️ **카드 선택에 따른 처치 효과**를 파악해야 함

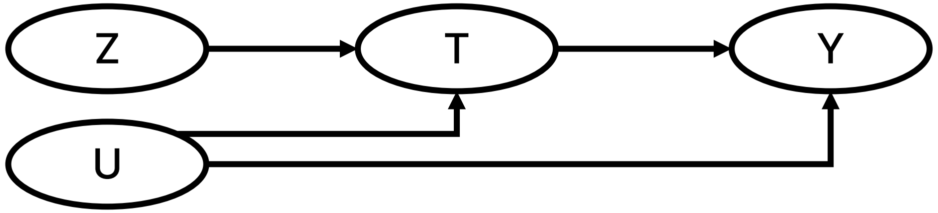
- 이 사례에서 프라임 카드 사용 자격이 없는 고객은 카드를 얻을 방법이 없음 ➡️ **단방향 불응** (one-sided non-compliance)이 발생
  - 순응 그룹의 수가 네 개에서 두 개가 됨 (항시 참여자는 순응자로, 반항자는 항시 불참자로 분류)

- 그렇다면 ITTE를 프라임 카드 효과의 대리 변수로 사용하면 어떨까? ➡️ **ITTE 추정값은 ATE 추정값보다 0에 가깝게 편향**될 수 있음
  - 실험군의 일부 대상이 실제로 처치 받지 않으므로 두 그룹 간의 차이가 실제보다 줄어듦

- 실험군과 대조군 간의 단순평균 차이는 어떨까? ➡️ 이 차이는 실제 효과보다 더 크게 나타남
  - 프라임 카드를 선택한 고객은 프라임 서비스와 관계 없이 더 많이 지출하기 때문에, 상향 편향이 발생

- ATE를 식별할 수 없고 ITTE는 ATE의 편향 추정값인 상황 ➡️ 추가적인 가정이 필요!

도구 변수 식별 가정



- **독립성**
  - Z와 T, Z와 Y 사이에 측정되지 않은 교란이 없다고 가정 (도구 변수가 무작위로 배정된 것처럼 작용)
  - 검증할 수는 없지만, 실험 설계를 통해 설득력을 높일 수 있음 (예시에서는 프라임 카드 사용을 무작위 배정했으므로 충족)

- **배제 제약** (exclusion restriction)
  - 처치 T를 통하지 않고 Z에서 Y로 가는 경로가 없다는 가정 ($Y_{z,t} = Y_t$), 즉, 도구 변수는 처치를 통해서만 결과에 영향을 미침
  - Z가 무작위로 배정되어도 다른 경로로 영향을 미칠 수 있으므로 검증하기 어려움

- **연관성** (relevance)
  - Z에서 T로 향하는 화살표가 존재함 ($E[T_1 - T_0] \neq 0$). 즉, 도구 변수가 처치 변수에 영향을 미쳐야 함.
  - 도구 변수가 처치 변수에 주는 영향을 추정할 수 있으므로 검증 가능함

- **단조성** (monotonicity)
  - 도구 변수의 영향을 받은 모든 대상의 처치 받을 확률이 단지 증가하거나 감소하는 것을 의미함 ($T_{i1} \geq T_{i0}$ 또는 그 반대).
  - 처치 받을 확률이 증가한다면 반항자가 없고, 감소한다면 순응자가 없다고 가정하는 것과 같음

도구 변수 식별 가정

- ITTE를 시작점으로 ATE를 추정해보자

$$ITTE = E[Y|Z=1] - E[Y|Z=0] = E\left[Y_{1,1}T_1 + Y_{1,0}(1-T_1)\middle|Z=1\right] - E\left[Y_{0,1}T_0 + Y_{0,0}(1-T_0)\middle|Z=0\right]$$

- 배제 제약 가정을 이용해 $Y_{z,t}$의 도구 변수 첨자를 제거

$$E[Y_1T_1 + Y_0(1-T_1)|Z=1] - E[Y_1T_0 + Y_0(1-T_0)|Z=0]$$

- 독립성 가정을 사용하여 두 기댓값을 하나로 묶고 단순화

$$E[Y_1T_1 + Y_0(1-T_1) - Y_1T_0 + Y_0(1-T_0)] = E[(Y_1-Y_0)(T_1-T_0)]$$

- 단조성 가정을 사용해, 기댓값을 $T_1 > T_0$와 $T_1 = T_0$인 경우로 확장

$$E[(Y_1-Y_0)(T_1-T_0)|T_1>T_0]P(T_1>T_0) + E[(Y_1-Y_0)(T_1-T_0)|T_1=T_0)P(T_1=T_0)$$

도구 변수 식별 가정

- $T_1 = T_0$이라면 $T_1 - T_0 = 0$이고, $T_1 > T_0$이면 $T_1 - T_0 = 1$이므로 ITTE는 다음과 같이 표현할 수 있음

$$E[Y|Z=1] - E[Y|Z=0] = E[Y_1 - Y_0|T_1 > T_0]P(T_1 > T_0)$$

- 즉, 도구 변수가 결과에 미치는 영향은 **순응자의 처치 효과에 순응률 (compliance rate)를 곱한 값**
  - 단조성 가정에 따라 $E[T_1 - T_0] = P(T_1 > T_0)$이며, 도구 변수가 무작위로 배정되었으므로 처치에 미치는 영향을 추정할 수 있음

- 종합하면, Z가 Y에 미치는 영향을 Z가 T에 미치는 영향인 순응률로 조정함으로써 순응자들에 대한 평균 처치효과를 식별할 수 있음

$$E[Y_1 - Y_0|T_1 > T_0] = \frac{E[Y|Z=1] - E[Y|Z=0]}{E[T|Z=1] - E[T|Z=0]}$$

- 이렇게 하면 ATE가 아닌 순응자에 대한 효과인 **국지적 평균 처치 효과** (LATE: Local Average Treatment Effect)만 식별 가능
  - 하지만, 경우에 따라 LATE를 아는 것만으로도 충분할 수 있음
  - 예시에서 LATE는 프라임 카드를 사용할 수 있을 때, 선택한 사람들에 대한 효과
  - 은행은 추가 금액 효과가 프라임 카드 비용을 상회하는지 알고 싶기 때문에 프라임 카드를 선택하지 않은 사람들을 신경 쓰지 않음

도구 변수 분석

- 도구 변수 분석은 두 단계로 진행됨

  1. 1단계 (first-stage) 회귀: 처치 변수를 도구 변수에 회귀하는 단계
  2. 2단계 또는 축약형 (reduced form) 회귀: 결과 변수를 도구 변수에 회귀하여 처치 의도 효과를 추정하는 단계

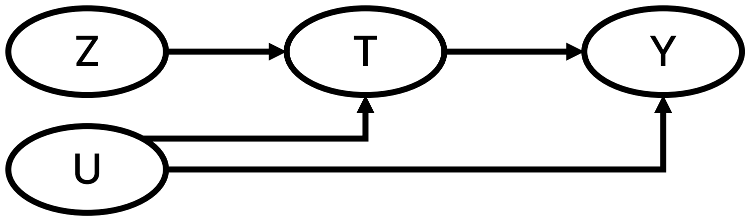
- 1단계 회귀의 매개 변수 추정값을 2단계 회귀의 매개 변수 추정값으로 나누어 LATE 추정값을 얻음

$$E[Y_1 - Y_0 | T_1 > T_0] = \frac{E[Y|Z=1] - E[Y|Z=0]}{E[T|Z=1] - E[T|Z=0]}$$

➡️ 2단계 회귀로 추정
➡️ 1단계 회귀로 추정

- 다만, 신뢰구간을 활용하지 않으면 이 추정값이 맞는지 확인하기 어려움

2단계 최소제곱법



- DAG에서 볼 수 있는 처치의 원인은 (1) 랜덤화된 도구 변수 Z와 (2) 교란 편향이 발생하는 U
- 2단계 최소제곱법의 1단계는 처치 변수를 도구 변수에 회귀하는 단계이며, 이는 경로 $Z \to T$를 추정하는 단계
    - 더 깊은 의미는, 1단계에서의 예측값인 $\hat{T}$를 편향되지 않은 처치 변수라고 볼 수 있음
- 2단계에서는 결과 변수를 예측값 $\hat{T}$에 회귀하여 이전과 같은 도구 변수 추정값을 얻게 됨
- 이 방법을 **2단계 최소제곱법** (2SLS: two-stage least squares)라고 하며, 이 방법이 유용한 이유는
    1. 표준오차를 적절하게 계산할 수 있고
    2. 더 많은 도구 변수와 공변량을 쉽게 추가할 수 있기 때문임

표준오차

- 2단계 예측의 잔차는 다음과 같이 정의할 수 있음

$$\hat{e}_{IV} = Y - \hat{\beta}_{IV} T$$

- 2단계 회귀의 .resid 메서드를 통해 얻을 수 있는 잔차는 $Y - \hat{\beta}_{IV}\hat{T}$ 이므로 동일한 잔차가 아니라는 것을 유의해야 함
    - 원하는 잔차는 예측된 처치 ($\hat{T}$)가 아닌 기존의 처치 (T)를 사용

- 해당 잔차를 사용하여 도구변수 추정값에 대한 표준 오차를 계산할 수 있음

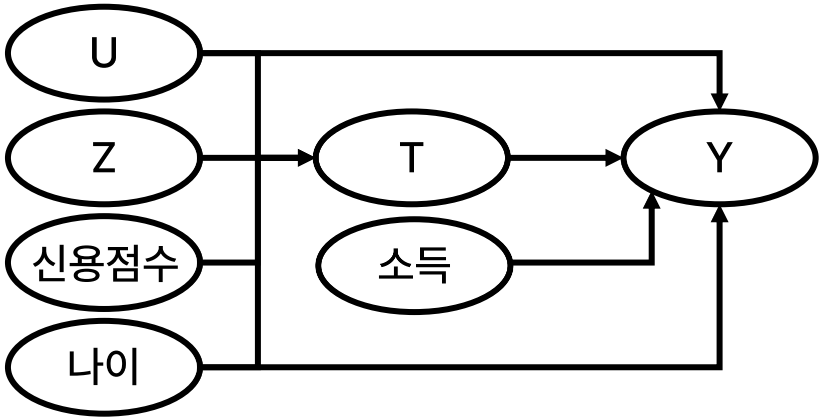
$$SE\left(\widehat{\beta_{IV}}\right) = \frac{\sigma(\hat{\epsilon}_{IV})}{\hat{\beta}_{z,1st}\sigma(Z)\sqrt{n}}$$

- σ는 표준편차, $\hat{\beta}_{z,1st}$는 1단계 회귀에서 얻은 예상 순응률을 의미함

    - 순응률이 100%인 경우, 표준오차는 OLS와 동일함
    - 그러나 불응이 존재하는 경우, 표준오차가 증가하게 됨 (순응률이 50%인 실험에 필요한 표본 크기는 100%인 경우의 4배)

통제변수와 도구변수 추가

- 도구변수 표준오차를 낮출 수 있는 방법 중 하나는 추가 공변량을 포함하는 방법



- 소득은 결과 예측력이 높지만 순응 여부를 예측하지 못하고, 신용점수는 순응 여부를 예측하지만 결과는 예측하지 못함
  - 나이는 순응과 결과 모두에 영향을 미치는 교란 요인

- 신용점수를 추가 도구 변수로 포함하면, LATE 매개변수의 표준오차를 크게 줄일 수 있음

- 하지만, 도구 변수 배정 메커니즘을 알지 못하면 배제 제약이 위배되는지 알기 어려움
  - 결과가 아닌 처치의 원인을 조건으로 두면 추정값의 분산이 증가

- 대부분의 경우 공변량이 순응과 결과 모두에 영향을 주며, 결과를 잘 예측할 통제 변수를 포함해서 분산을 줄일 수 있음
  - 이 예시에서는 소득을 추가 통제변수로 포함하면 표준오차를 상당히 줄일 수 있음

통제변수와 도구변수 추가

- Linearsmodel 라이브러리를 사용하지 않고 2SLS를 직접 구현하면 동일한 추정값을 얻을 수 있지만 표준오차는 달라짐
  - 처치 행렬과 도구변수 행렬에 추가 공변량을 포함시켜 행렬을 통해 도구변수의 잔차와 분산을 구할 수 있음

$$\widehat{Var}\left(\hat{\beta}_{IV}\right) = \sigma^2(\hat{e}_{iv})diag(\left(\hat{X}'\hat{X}\right)^{-1})$$

- 행렬 표기법 때문에 분산 공식을 해석하기 어렵지만, 추가 공변량 없이도 이전과 더 유사한 값을 구할 수 있음

$$SE\left(\hat{\beta}_{IV}\right) \approx \frac{\sigma(\hat{e}_{iv})}{\sigma\left(\tilde{T}\right)\sqrt{nR_{1st}^2}}$$

- $\tilde{T}$는 처치를 추가 공변량 (도구변수 제외)에 회귀한 처치의 잔차이며, R_{1st}^2는 1단계 회귀의 R^2

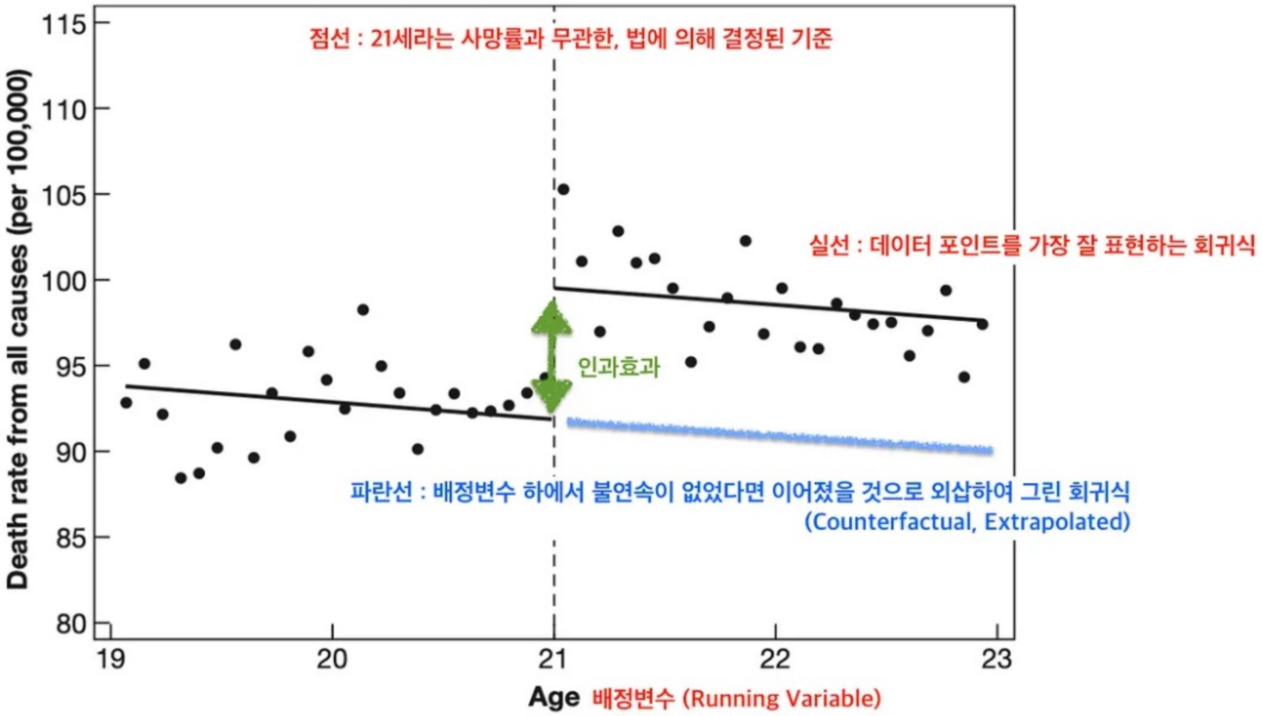
- 위 시에서 표본 크기를 늘리는 방법 외에 표준오차를 줄이는 세 가지 방법이 있음을 알 수 있음

  1. 1단계 회귀의 R^2 증가시키기. 이는 순응을 잘 예측하는 강력한 도구변수를 찾아서 해결할 수 있음 (배제 제약을 만족해야 함)
  2. 처치 T에 대한 예측력이 높은 변수를 제거하여 $\sigma\left(\tilde{T}\right)$ 증가시키기
  3. 결과의 예측력이 높은 변수를 찾아서 2단계 회귀의 잔차 줄이기

- 실전에서 도구변수를 찾는 것은 매우 어렵고, 2번째 방법 역시 한계가 있으므로 마지막 방법이 가장 현실적

불연속 설계 (Regression Discontinuity Design)

- 처치 배정에 인위적인 불연속성을 활용하여 처치 효과를 식별하는 방법
- 미국의 법적 최소 음주 연령 (MLDA) 제도가 사망률에 주는 영향을 알고 싶을 때, 다음과 같은 설계를 통해 인과 효과를 분석할 수 있음
  - 처치 (T): 술을 마실 수 있도록 합법적으로 허용하는 것 (21세를 기준으로 1과 0으로 나뉨)
  - 배정변수 (Running Variable): 처치를 결정하는 변수 (나이 또는 출생년월을 의미함)
  - 결과변수: 나이별 사망률



Source : https://medium.com/bondata/regression-discontinuity-bba1c4d7a08c

RDD와 이중차분법의 차이점

- 이중차분법: 대조군으로 여겨지는 그룹의 처치 전후 데이터가 존재해, 대조군과의 차이를 기반으로 인과효과를 추정

- RDD와 이중차분법의 차이점

  - RDD에는 배정변수라는 개념이 추가됨
    - 배정변수 전후 그룹에는 차이가 없을 것이라는 가정 (e.g. 20~22세 그룹의 인구통계학적 특성은 유사할 것이다)
    - 다만, 얼마나 전후까지 차이가 없을 것이라는 질문이 파생됨 ➡️ Bandwidth 설정

  - RDD에는 대조군으로 여겨지는 그룹의 처치 전후 데이터가 존재하지 않음
    - 배정변수 전후 데이터 포인트를 기반으로 추세를 학습함
    - 따라서 데이터 포인트를 잘 설명할 수 있는 모형이 무엇일까? 라는 질문이 파생됨 ➡️ Parametric 모델 설정

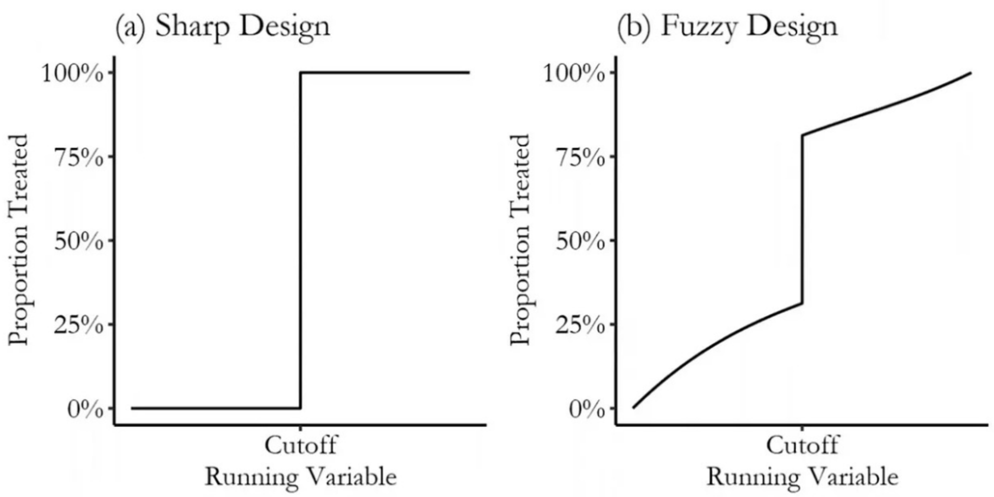
- 둘 다 적용할 수 있는 상황이라면?

  - 이중차분법의 가정 (Parallel Trend Assumption)은 인과추론 기법 중 가장 검증이 쉬운 가정
  - RDD의 가정은 검증하기 어려움 (임계치 전후로 특성이 유사해야 함)
  - 따라서, 둘 다 적용할 수 있는 상황에서는 가정을 충족시키거나 검증하기 용이한 이중차분법이 우위에 있음

Source : https://medium.com/bondata/regression-discontinuity-bba1c4d7a08c

RDD의 종류

- 앞선 예시처럼 임계치를 기점으로 처치 여부가 깔끔하게 결정되는 경우는 많지 않음
    - 계단형 회귀 불연속 (Sharp RD): 배정변수가 임계치를 통과함에 따라 처치 여부가 0에서 1로 변함
    - 경사형 회귀 불연속 (Fuzzy RD): 배정변수가 임계치를 통과함에 따라 처치 확률이 급격히 변함 ➡️ Imperfect Compliance 발생



- 임계값을 초과하더라도, 처치 확률이 1보다 작으면 관측한 결과가 실제 잠재적 결과보다 작아짐
    - 마찬가지로, 임계값 아래에서 관측한 결과는 실제 잠재적 결과보다 높음
    - 따라서 임계값에서의 처치효과가 실제보다 작게 보이며, **도구변수를 통해 이를 보정해야 함**

Source : https://medium.com/bondata/regression-discontinuity-bba1c4d7a08c

RDD와 신용카드 예시

- 은행에서 모든 고객에게 신용카드를 제공하지만, 계좌 잔고가 $5,000 이하인 고객에게는 수수료를 부과한다고 가정해보자

- 도구변수 추정값의 분자에 해당하는 처치 의도 효과 (ITTE)를 다음과 같은 과정을 통해 추정

  1. 배정변수인 잔고에서 5,000을 빼 임계값을 0으로 옮김
  2. 임계값 초과 여부를 나타내는 더미변수와 조정된 배정변수 R을 사용해 결과변수에 대한 회귀분석을 진행함
     $$y_i = \beta_0 + \beta_1 r_i + \beta_2 \mathbf{1}(r_i > 0) + \beta_3 \mathbf{1}(r_i > 0) r_i$$
  3. 임계값 초과 여부와 관련된 매개변수 추정값 $\hat{\beta}_2$를 처치 의도 효과로 해석할 수 있음

- 순응률이 100%가 아니므로 ITTE를 순응률로 나누어야 함 ➡️ 이전 과정에서 결과변수를 처치변수로 대체해 반복하면 됨

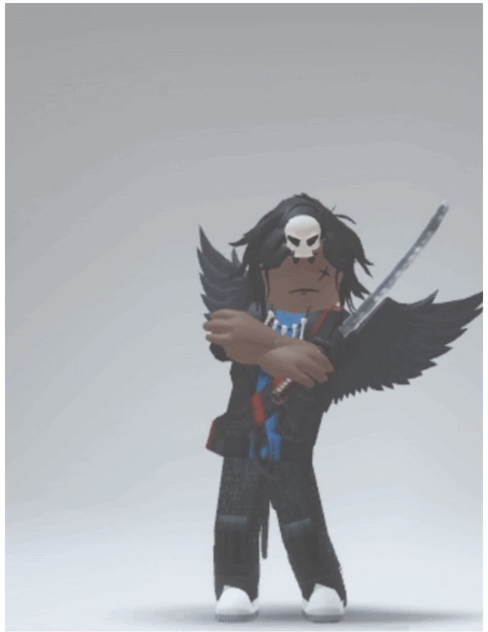
- 불연속 설계의 또 다른 잠재적 문제는, 대상이 배정변수를 조작해 스스로를 실험군에 포함시킬 수 있다는 것

  - 신용카드 예시의 경우, 프라임 카드를 무료로 받기 위해 잔고를 $5,000 이상으로 조정할 수 있음
  - 임계값 주변의 밀도를 그래프로 나타내 이런 일이 발생하는지 시각적으로 확인할 수 있음
  - 실험 대상들이 스스로를 실험군에 포함시킨다면, 임계값에서 밀도가 증가하는 현상을 예상할 수 있음

2. 예시

Causal Inference using Instrumental Variables (Roblox)

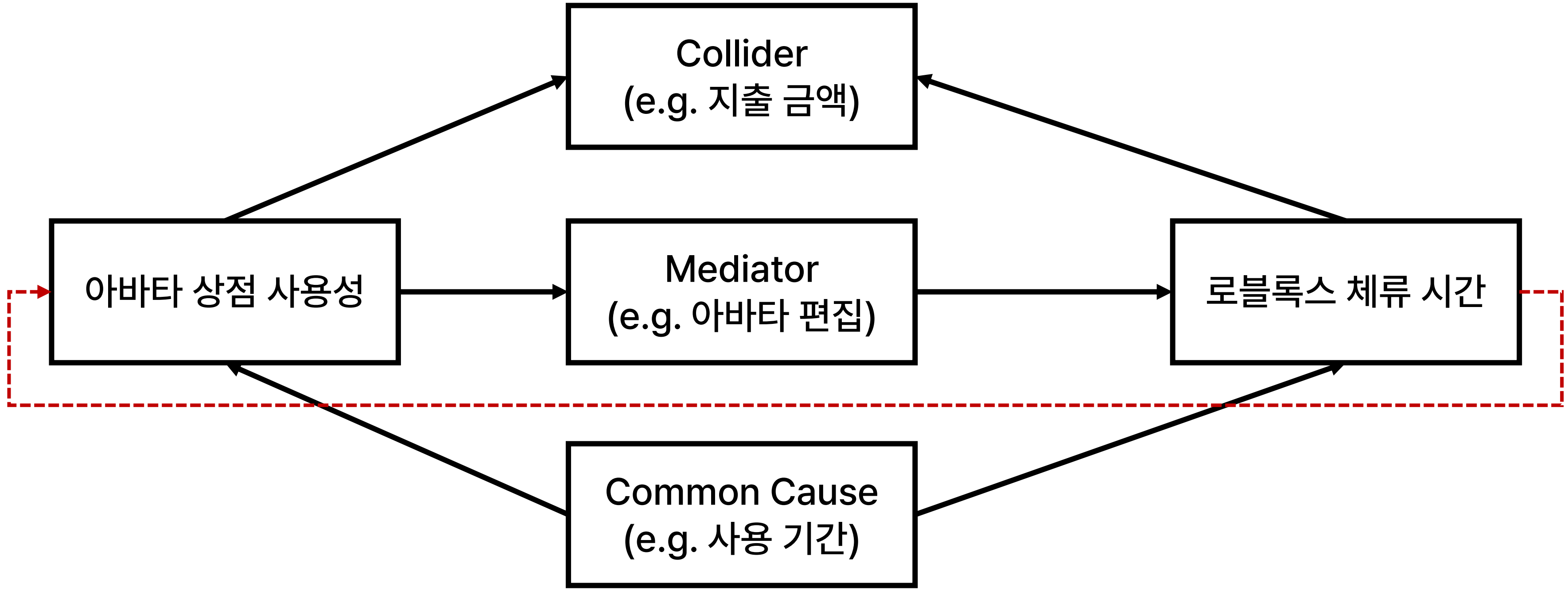
- 로블록스에서는 아바타 상점에서 구매한 아이템으로 사용자의 아바타를 꾸밀 수 있음
- 아바타 상점 사용성이 로블록스 체류 시간에 미치는 영향을 실험을 통해 파악하기에는 어려움이 있음
  - 아바타 상점은 매우 중요한 기능이므로 일부 사용자에게 아바타 상점 기능을 제공하지 않는 것은 불가능함
  - 아바타 상점은 판매자와 구매자가 상호작용하는 공간이므로 일부에게 기능을 제공하지 않는 것이 다른 사용자에게도 영향을 미침



Source : https://blog.roblox.com/2021/09/causal-inference-using-instrumental-variables/

Causal Inference using Instrumental Variables (Roblox)

- 또한, 여러 교란 요인과 체류 시간이 역으로 상점 사용성에 미치는 영향 때문에 실험이 아닌 환경에서 영향을 분석하는 것 역시 어려움



- 이중차분법, PSM 등의 인과추론 방법이 있지만 교란 변수 또는 시간에 따라 변하는 교란 변수에 취약하다는 단점이 있음

Source : https://blog.roblox.com/2021/09/causal-inference-using-instrumental-variables/

Causal Inference using Instrumental Variables (Roblox)

- 도구 변수가 이 문제의 해결 방법이 될 수 있지만, 가정을 모두 충족시키는 변수를 찾는 것 역시 어려움

- 2020년에 진행한 아바타 상점 추천 A/B 테스트의 실험군 배정 여부를 도구 변수로 사용함

  - 어떤 그룹에 속했는지 여부가 아바타 상점 사용성에 큰 영향을 미치는 것을 관측함 (연관성)
  - 해당 실험은 아바타 상점 내 변화를 평가하기 위해 설계되었고 이외에는 아무 변화가 없었기 때문에 체류 시간의 변화는 오롯이 아바타 상점 사용성의 변화로 인한 것이라고 가정할 수 있음 (배제 제약)

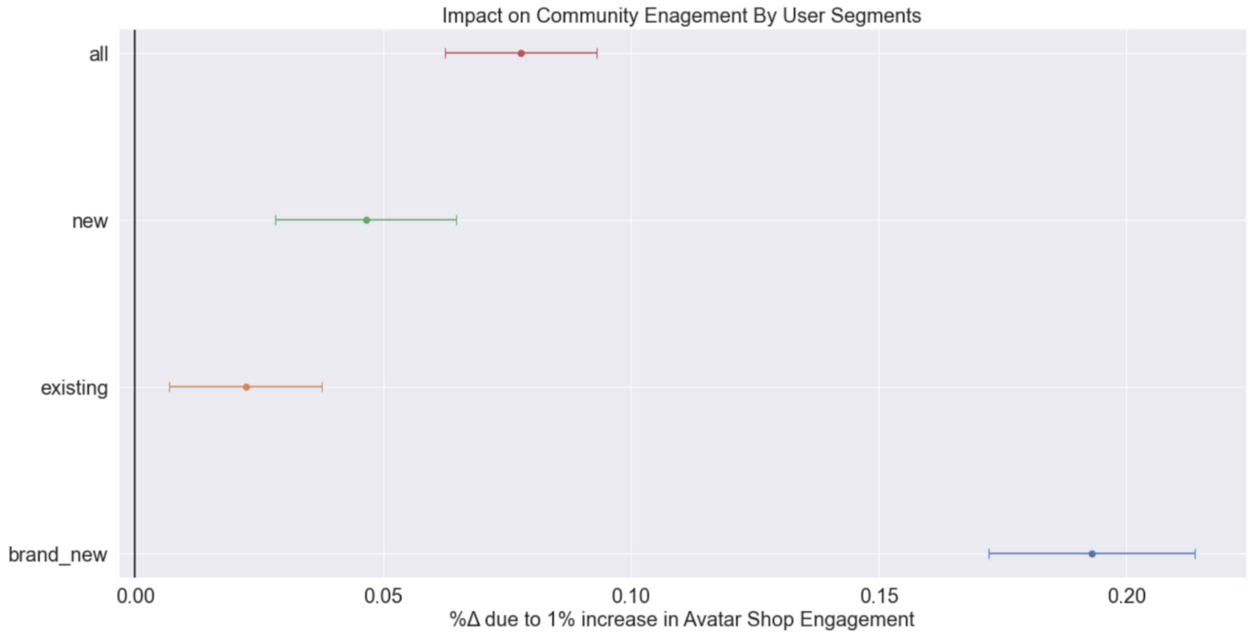
Source : https://blog.roblox.com/2021/09/causal-inference-using-instrumental-variables/

Causal Inference using Instrumental Variables (Roblox)

- 도구변수 추정 결과 아바타 상점 사용성과 로블록스 체류 시간 사이에 통계적으로 유의한 양의 관계가 있음을 확인함
  - 아바타 상점 사용성의 1% 증가가 0.08%의 체류 시간 증가로 이어짐 (p-value < 0.0001)

- 흥미로운 것은, 사용자 segment에 따라 영향이 다르게 나타남
  - 가입한 지 일주일이 되지 않은 brand new 사용자에게 미치는 영향이 가장 컸음

- 이는 ATE가 아닌 LATE를 추정하는 도구변수 분석 방법의 한계점을 보여주기도 함
  - 추정된 인과효과는 도구변수에 영향을 받은 그룹에 한정적이며, 전체 사용자에게 일반화하기는 어려움



Source : https://blog.roblox.com/2021/09/causal-inference-using-instrumental-variables/

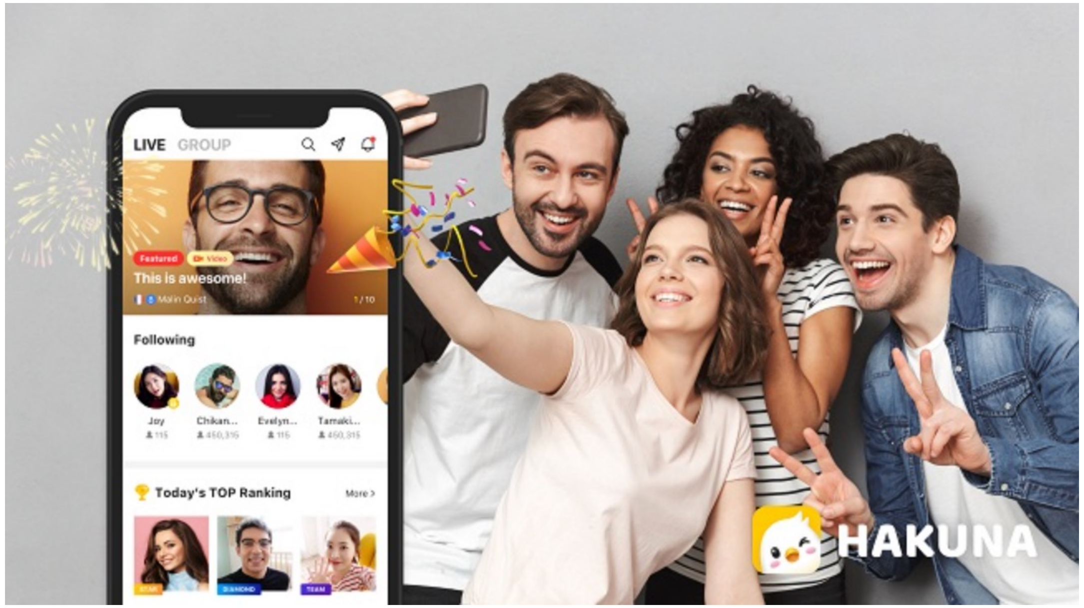
회귀 단절 모형을 활용한 전후 비교 분석 (하이퍼커넥트)

- RDiT (Regression Discontiunity in Time)은 독립변수가 시간인 RDD로, 일반적인 RDD와 다음과 같은 차이점이 있음
    - 실험 설계: 예를 들어 대기 오염 물질을 배출하는 공장을 규제하는 정책의 효과를 분석할 때 RDD가 1000mw보다 큰 규모와 작은 규모의 공장을 비교한다면, RDiT는 1000mw보다 큰 공장에 한해 정책 도입 전후를 비교 (동일 집단의 서로 다른 시점을 비교)
    - 비교 집단 구성: RDD의 경우 임계치 (1000mw) 부근의 값끼리 비교한다면, RDiT는 임계점에 근접한 날짜끼리 비교하기 어려움 (4월 3일은 평일이고 5일이 공휴일이라면 임계점 부근이라 해도 비교하기에는 어려움이 있음)
    - 샘플 사이즈: RDD에서는 샘플 사이즈를 가능한 만큼 늘릴 수 있지만 (더 많은 공장을 관찰), RDiT는 관찰 기간이 넓어지므로 수많은 변수가 생기게 됨

Source : https://hyperconnect.github.io/2021/06/07/regression-discontinuity-in-time.html

회귀 단절 모형을 활용한 전후 비교 분석 (하이퍼커넥트)

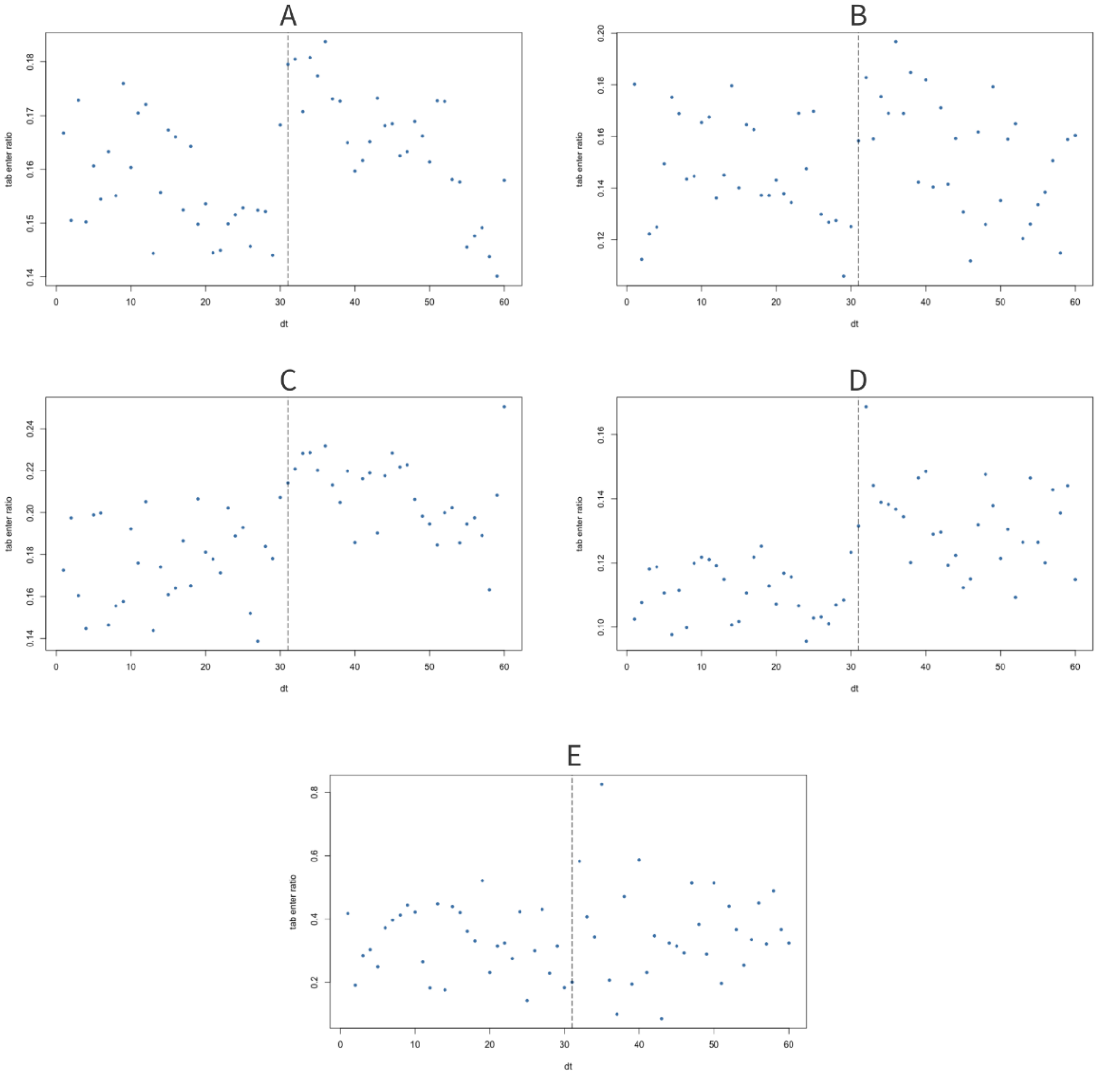
- 하이퍼커넥트의 라이브 스트리밍 서비스인 하쿠나 라이브에는 여러 사람이 함께 방송할 수 있는 라운지 (Lounge)가 있음
    - 2021/02/24에 라운지 진입을 늘리기 위한 팝업을 배포함
    - RDiT를 활용한 전후 비교의 예시로 팝업 배포 전후 탭 진입율이 유의미하게 변화했는지 분석
- 각 국가별로 임계 시점 전후 30일치 데이터를 확보하고, 해당 기간 팝업 배포 외 진입율에 영향을 미친 어떠한 이벤트도 없음을 확인함



Source : https://hyperconnect.github.io/2021/06/07/regression-discontinuity-in-time.html

회귀 단절 모형을 활용한 전후 비교 분석 (하이퍼커넥트)

- 임계 시점 전후에 선형적 추세가 관찰됨
- A, C, D 국가의 경우 임계 시점 이후 진입율이 상승한 부분이 명확하게 관찰됨
- 반면, B와 E 국가는 크게 달라지지 않음



Source : https://hyperconnect.github.io/2021/06/07/regression-discontinuity-in-time.html

회귀 단절 모형을 활용한 전후 비교 분석 (하이퍼커넥트)

- R의 rddtools 패키지를 활용해 데이터를 RDD 모형에 fitting함
  - 임계 시점에 해당하는 숫자를 cutpoint로 입력
  - 전후 데이터의 기울기가 유사한 편이므로 slope = "same"으로 지정
- 팝업 배포로 인해 A국가의 탭 진입율이 3%p 증가함
  - E국가를 제외한 다른 국가 모두 탭 진입율이 유의하게 상승함

```
Call:
lm(formula = y ~ ., data = dat_step1, weights = weights)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0167380 -0.0047000 -0.0003103  0.0048069  0.0220548

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.1454416  0.0024312  59.823  < 2e-16 ***
D            0.0303626  0.0042636   7.121 2.00e-09 ***
x           -0.0007516  0.0001231  -6.106 9.65e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.008253 on 57 degrees of freedom
Multiple R-squared:  0.4709,	Adjusted R-squared:  0.4523
F-statistic: 25.36 on 2 and 57 DF,  p-value: 1.321e-08
```

```r
# import library
install.package("rddtools")
library(rddtools)

# import raw data
raw = read.csv("file-path", header=T)

# construct RD data
region_A = raw[,c(1,2)]
data <- rdd_data(y=region_A[,2],
                 x=region_A[,1],
                 data=region_A,
                 cutpoint=31)

# estimate the sharp RD model
rdd_mod <- rdd_reg_lm(rdd_object=data,
                      slope="same")
summary(rdd_mod)
```

Source : https://hyperconnect.github.io/2021/06/07/regression-discontinuity-in-time.html

감사합니다
Q&A