가짜연구소

Causal-Lab

Part 1.인과추론의 기초

인과추론 소개

2025.03.09 박이삭

발표자 소개

박이삭 (Isaac Park)

HYBE IM | 데이터 분석가
- 2018년 ~ 2023년 라인게임즈
- 2023년 ~ 현재

가짜연구소 | 10기 빌더

2025년 목표: 오픈 소스에 기여 하자

Constens

Causal-Lab

인과추론 소개

- 개념
- 목적
- 머신러닝과 인과추론
- 연관관계 & 인과관계 (+편향)
- 가정
- 요약

인과추론 개념

Causal-Lab

뭘까?

인과추론 개념

Causal-Lab

원인과 결과를 가장 과학적으로 접근하는 학문

과학?
- 가설을 세우고 실험을 하는 학문

즉, 인과추론은 “원인”과 “결과”를 가장 잘 설명하는 학문이다
어떻게? 가설과 실험으로 검증하면서!

인과추론 개념

인과추론 개념

Causal-Lab

원인: 무슨일이 생겼다

결과: 나간다

원인: 무슨일이 생겼다

결과: 잔다

인과추론 개념

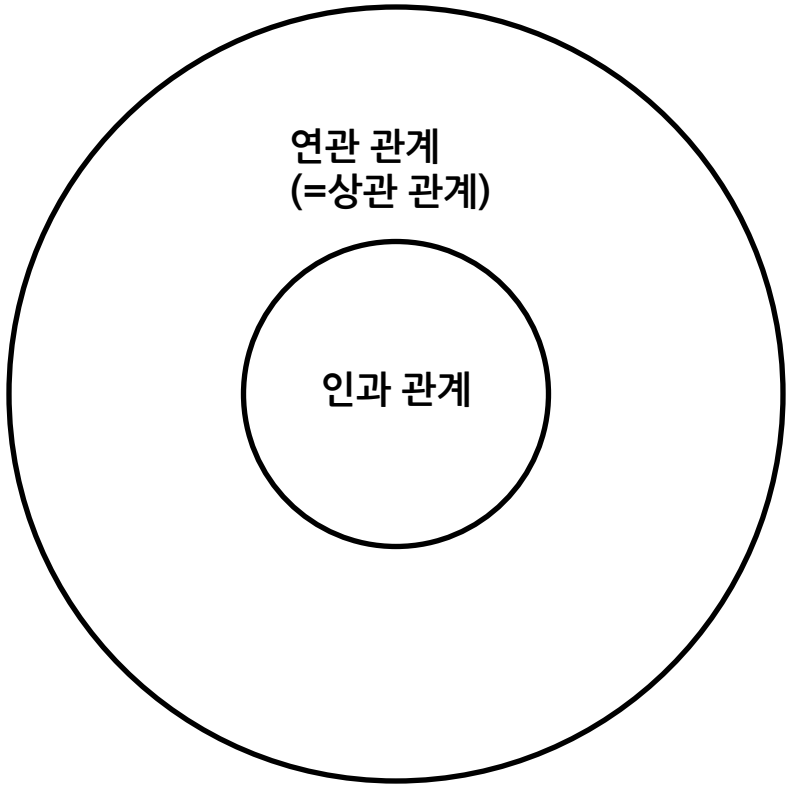
출처: Link

인과추론 개념

Causal-Lab

연관관계는 때로 인과관계가 될 수도 있다 (p35)

인과추론은 이 둘을 구분하고 왜 다른지 이해하는 **과학**이다. (p36)



인과추론 개념

교회가 많아질수록 범죄율이 증가한다

운동을 하면 몸이 건강해진다

인과추론 개념

Causal-Lab

연관관계

교회가 많아질수록 범죄율이 증가한다

인과관계

원인

결과

운동을 하면 몸이 건강해진다

인과추론의 목적

Causal-Lab

인과추론의 목적

Causal-Lab

어느날 지표를 확인해보니, Level이 높을수록 ARPPU(=결제 유저당 평균 매출)이 높다는 것을 확인

게임 속 레벨이 높을수록 과금을 많이 한다.?

만약, 위 가설이 맞다면

우리회사가 돈을 벌려면 유저들 레벨을 올려주어야 하는구나! 라는 결론을 내릴 수 있습니다.

인과추론의 목적

Causal-Lab

인과추론을 해보자면...

로그인 데이터를 보니, 낮은 레벨의 유저들은 이탈을 많이하고
레벨이 높을수록 로그인을 더 자주하고 있다면?

과금할 확률이 동일하다면, 더 자주 접속할수록 과금을 많이 하겠죠?
그리고 자주 접속할수록 레벨도 높아짐

즉, 우리회사가 돈을 벌려면 유저가 더 자주 접속하게끔 만들어야 한다

이런 질문들은 비즈니스나 삶에서 무언가를 변화시켜 더 나은 결론을 얻고자 하는 욕구에서 비롯됩니다. (p37)

머신러닝과 인과추론

Causal-Lab

인과추론은 'what if' 유형의 질문(가설)을 다루고 있습니다.

데이터 설계를 잘하면 인과추론에도 머신러닝을 이용할 수 있지만
대부분의 경우에는 그대로 사용할 수 없습니다.

왜냐하면, 머신러닝 $X \propto Y$ 연관관계를 찾기 때문

X		Y
성수기 구분	숙박 요금	예약 여부
0	10만원	1
0	11만원	0
0	9만원	1
0	8만원	0
1	20만원	1
1	19만원	1
1	22만원	1
1	19만원	1

P38,호텔 성수기 예시

→ 7장 메타러너에서 머신러닝을 통한 인과추론 방법을 배울 예정입니다.

연관관계와 인과관계

Causal-Lab

A (재미존)

B (오즈토이)

본격적으로...

“기업의 주요 목표는 당연히 매출의 증대” 입니다.

언제 얼마만큼의 가격 할인을 하면 추가 이익이 발생할 것인가?

가격할인이 판매량에 미치는 영향을 파악해보자!

연관관계와 인과관계

판매량 ← f(가격할인)

가설: 가격 할인을 하면 매출이 증가 하는가?

연관관계와 인과관계

Causal-Lab

판매량 ← f(가격할인)

가설: 가격 할인을 하면 매출이 증가 하는가?

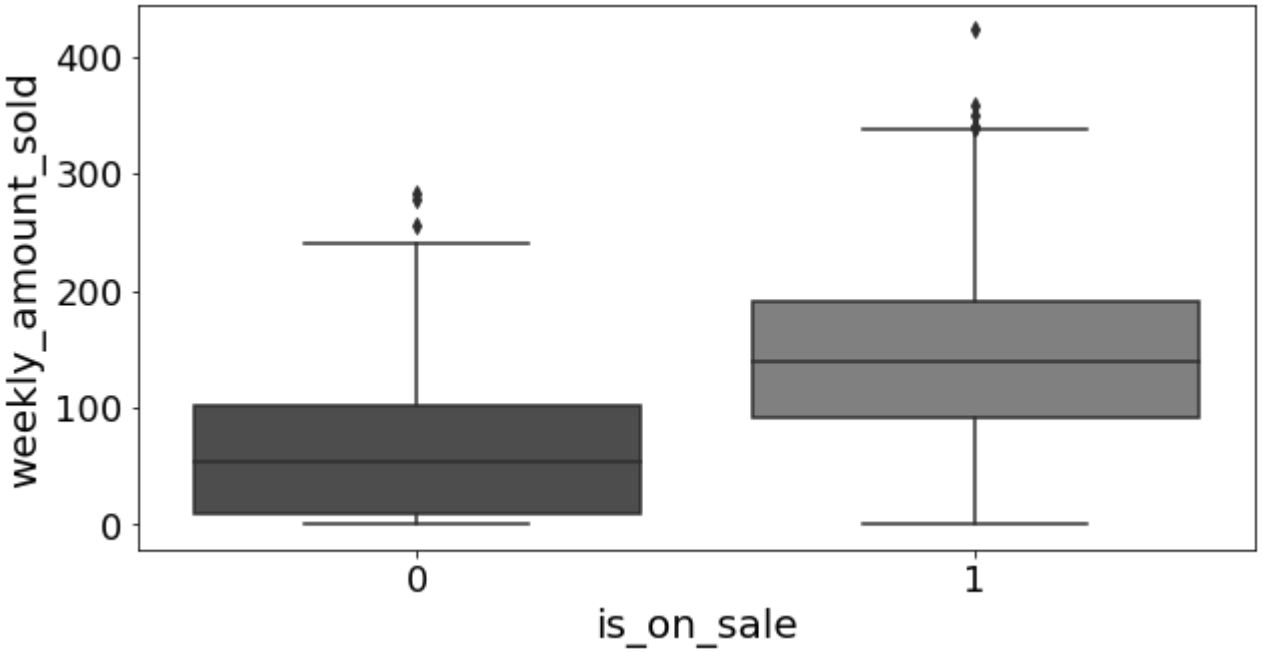
“원인” : 가격 할인	T_i	$T_{재미존} = 0 \; or \; 1$
“결과” : 매출 증가	Y_i	$Y_{0재미존} = 30$ $Y_{1재미존} = 40$

연관관계와 인과관계

	X_1	X_2	T	Y
상점	크리스마스까지 남은 기간(w)	연간 판매량	할인 진행여부	주간 판매량
재미존	3	1670	1	21
재미존	2	1670	1	18
재미존	1	1670	1	14
재미존	0	1670	0	10
오즈토이	3	1423	0	10
오즈토이	2	1423	0	5
오즈토이	1	1423	0	4
오즈토이	0	1423	0	7
...	...	...	...	...

P40 데이터 각색

연관관계와 인과관계



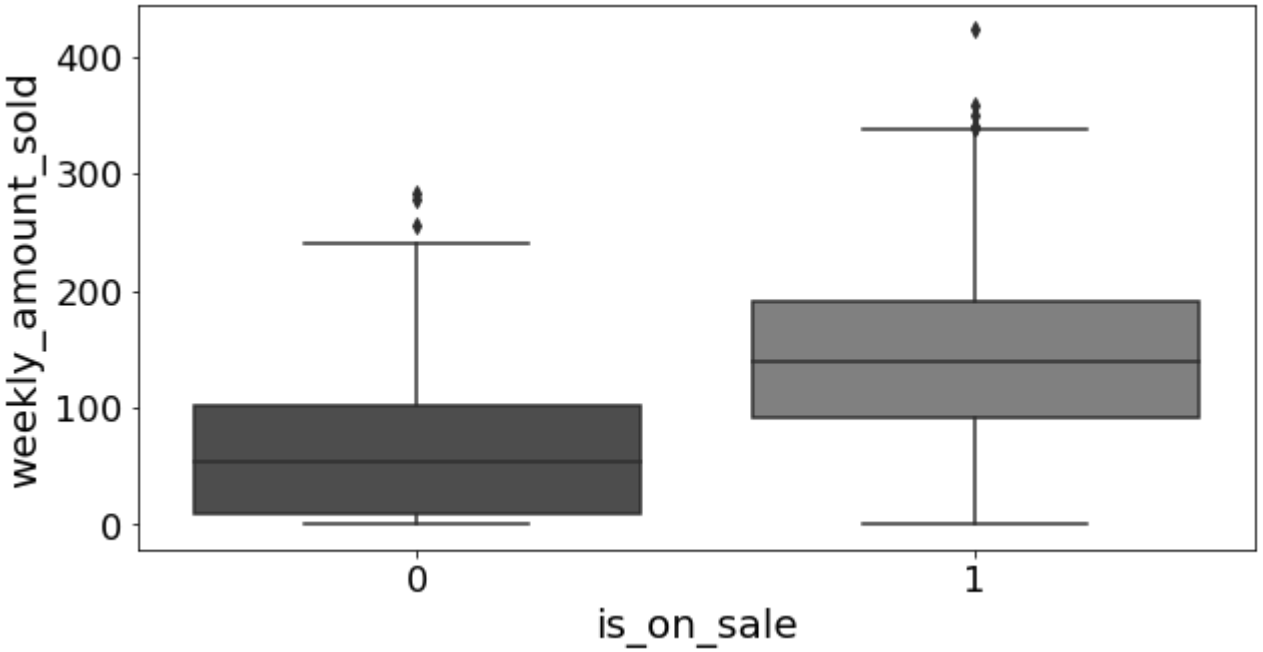
[그림 1-1]

연관관계와 인과관계

영향을 줄 수 있는 변수

- 연간 판매량 (=리뷰 수)
- 할인 기간

정확하게 측정해볼 수는 없을까?



[그림 1-1]

연관관계와 인과관계

영향을 줄 수 있는 변수

- 연간 판매량 (=리뷰 수)
- 할인 기간

정확하게 측정해볼 수는 없을까?

$$\tau_i = Y_{1i} - Y_{0i}$$

	X_1	X_2	T	Y0	Y1	diff
상점	크리스마스까지 남은 기간(w)	연간 판매량	할인 진행여부	주간 판매량	주간 판매량	
재미존	3	1670	1	10	21	11
재미존	2	1670	1	12	18	6
재미존	1	1670	1	13	14	1
재미존	0	1670	0	10	15	5
오즈토이	3	1423	0	10	13	3
오즈토이	2	1423	0	5	7	2
오즈토이	1	1423	0	4	9	5
오즈토이	0	1423	0	7	10	3
...	...	...	...	...	...	

연관관계와 인과관계

영향을 줄 수 있는 변수

- 연간 판매량 (=리뷰 수)
- 할인 기간

정확하게 측정해볼 수는 없을까?

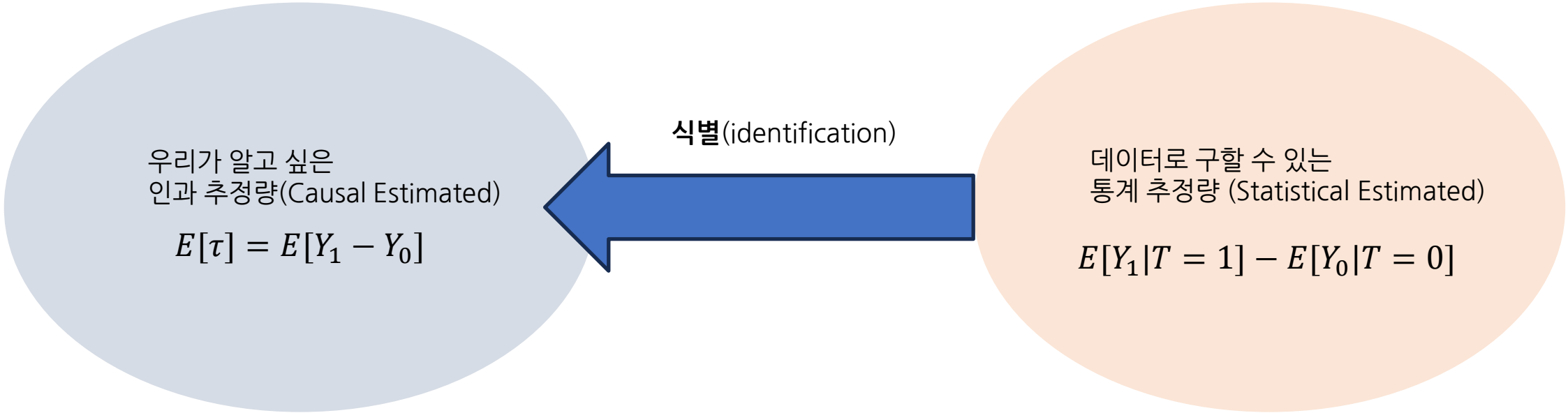
$$\tau_i = Y_{1i} - Y_{0i}$$

	X_1	X_2	T	Y0	Y1	diff
상점	크리스마스까지 남은 기간(w)	연간 판매량	할인 진행여부	주간 판매량	주간 판매량	
재미존	3	1670	1	반사실	21	?
재미존	2	1670	1		18	
재미존	1	1670	1		14	
재미존	0	1670	0	10	Counterfactual	
오즈토이	3	1423	0	10		
오즈토이	2	1423	0	5		
오즈토이	1	1423	0	4		
오즈토이	0	1423	0	7		
...	...	...	...	...	...	

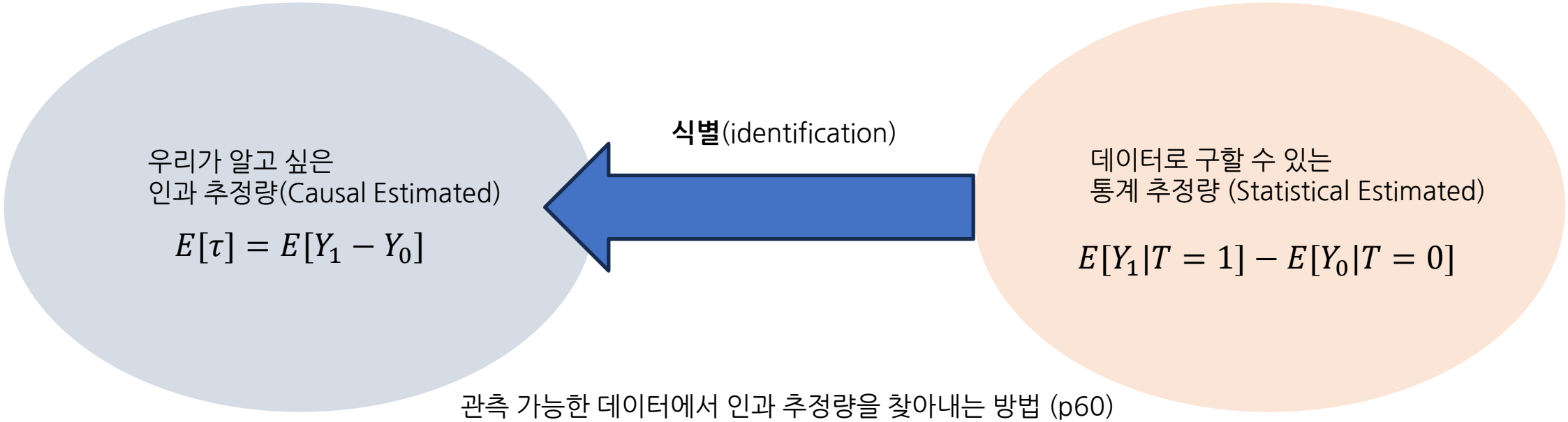
인과 추론의 근본적인 문제

연관관계와 인과관계

그림 - [실무로 통하는 인과추론] 1-1. 인과추론 기초



연관관계와 인과관계



관측 가능한 데이터에서 인과 추정량을 찾아내는 방법 (p60)

$$E[Y|T = 1] - E[Y|T = 0] = E[Y_1 - Y_0|T = 1] + \{E[Y_0|T = 1] - E[Y_0|T = 0]\}$$

인과효과 편향

연관관계와 인과관계

Causal-Lab

인과추론에서 식별에 대해 자세히 설명해줘

인과추론에서 '식별(identification)'은 관측된 데이터로부터 우리가 원하는 인과 효과를 정확하게 추정할 수 있는지 여부를 판단하고, 가능하다면 그 방법을 찾는 과정을 의미합니다. 즉, 데이터만 가지고 '원인'이 '결과'에 미치는 순수한 영향을 밝혀낼 수 있는지를 확인하는 것입니다.

식별의 중요성:

- 정확한 인과관계 파악: 상관관계만으로는 인과관계를 알 수 없습니다. 식별은 혼란 변수의 영향을 제거하고 진정한 인과 효과를 추정하는 데 필수적입니다.
- 정책 결정 및 실용적 응용: 식별 없이는 특정 정책이나 개입이 원하는 결과를 가져올지 확신할 수 없습니다. 따라서 식별은 실제 문제 해결을 위한 과학적 근거를 제공합니다.

식별의 핵심 개념:

- 혼란 변수(confounding variable): 원인과 결과 모두에 영향을 미치는 변수로, 인과관계 추론을 방해합니다.
- 반사실(counterfactual): 실제로 발생하지 않은 상황에 대한 가상적인 결과입니다. 예를 들어, 특정 치료를 받은 환자의 결과와 받지 않았을 경우의 결과를 비교하는 것입니다.
- 식별 전략(identification strategy): 혼란 변수를 통제하고 반사실을 추론하여 인과 효과를 추정하는 방법입니다.

우리
인과

Estimated)

$E[$ $|T = 0]$

$E[Y|T =$ $T = 0]\}$

연관관계와 인과관계

Causal-Lab

우리가 알
인과 추정

$E[\tau]$ =

주요 식별 전략:

- 무작위 대조 실험(randomized controlled trial, RCT): 처치를 무작위로 할당하여 혼란 변수의 영향을 제거합니다. 가장 이상적인 방법이지만 윤리적, 현실적 제약이 있을 수 있습니다.
- 도구 변수(instrumental variable, IV): 원인 변수에만 영향을 미치고 결과 변수에 직접적인 영향을 미치지 않는 변수를 사용하여 인과 효과를 추정합니다.
- 회귀 단절 설계(regression discontinuity design, RDD): 특정 기준값을 기준으로 처치가 결정될 때, 기준값 근처에서 처치 효과를 추정합니다.
- 차분 추정법(difference-in-differences, DID): 처치 집단과 통제 집단을 시간에 따라 비교하여 처치 효과를 추정합니다.
- 뒷문 조절(backdoor adjustment): 인과 그래프를 기반으로 혼란 변수를 식별하고 통제하여 인과 효과를 추정합니다.

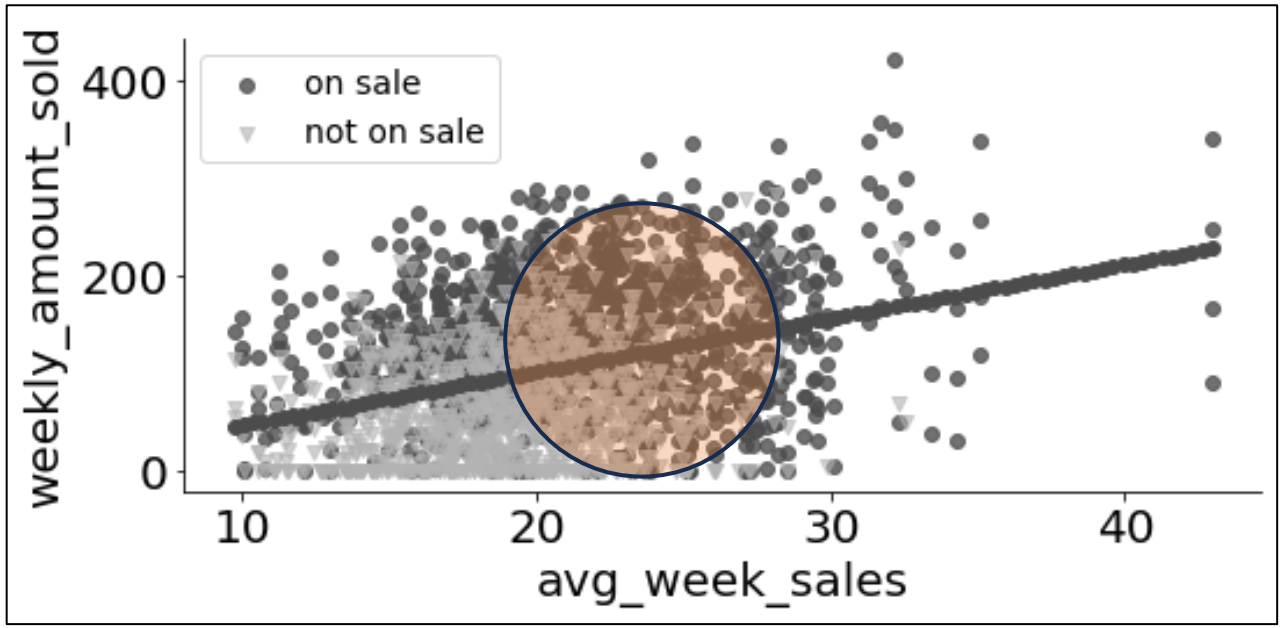
는
tical Estimated)

$[Y_0|T=0]$

$$E[Y|T=1] - E[Y|T=0] = E[Y_1 - Y_0|T=1] + \{E[Y_0|T=1] - E[Y_0|T=0]\}$$

인과효과 편향

가정(assumption)



1. 실험군과 대조군이 서로 교환 가능해야 한다.

→ 왜냐하면, 할인 유무 이외에 다른 조건이 비슷하다면(즉, 교환이 가능하다면) 두 그룹은 할인 때문에 차이가 발생.

최대한 성질이 비슷한 데이터를 모아서 실험을 진행(= 교환 가능해야 한다)

어떻게? 랜덤으로 진행하면 된다.

가정(assumption)

Causal-Lab

1. 실험군과 대조군이 서로 교환 가능해야 한다.
2. T(처치)와 잠재적 결과(Y)는 서로 독립이다

	X_1	X_2	T	Y0	Y1	diff
상점	크리스마스까지 남은 기간(w)	연간 판매량	할인 진행여부	주간 판매량	주간 판매량	
재미존	3	1670	1	10	21	11
재미존	2	1670	1	12	18	6
재미존	1	1670	1	13	14	1
재미존	0	1670	0	10	15	5
오즈토이	3	1423	0	10	13	3
오즈토이	2	1423	0	5	7	2
오즈토이	1	1423	0	4	9	5
오즈토이	0	1423	0	7	10	3
...	...	...	...	...	...	

T = 1 일 때 Y랑, T=0일 때 Y는 서로 영향을 주지 않는다

가정(assumption)

1. 실험군과 대조군이 서로 교환 가능해야 한다.
2. T(처치)와 잠재적 결과(Y)는 서로 독립이다
3. T(처치)와 잠재적 결과(Y)는 일치성(consistency)을 만족해야 한다.

일치?
- 처치를 여러 번 했음에도 일부만 고려한다면 일치성 위반

Ex)실험에 참가한 다른 기업들은 3주전 부터 10% 할인을 하는데
누군가는 욕심을 내서 화수목 요일에는 15% 할인을 한다면...

지정된 처치(T = 10% 할인) 이외에 숨겨진 처치가 있었음
이런 sample은 분석에서 제외되어야 함

가정(assumption)

Causal-Lab

1. 실험군과 대조군이 서로 교환 가능해야 한다.
2. T(처치)와 잠재적 결과(Y)는 서로 독립이다
3. T(처치)와 잠재적 결과(Y)는 일치성(consistency)을 만족해야 한다.
4. 상호 간섭 없음(SUTVA) = 파급 효과 없음

   A가 할인을 한다고 해서 B가 영향을 받지 않음

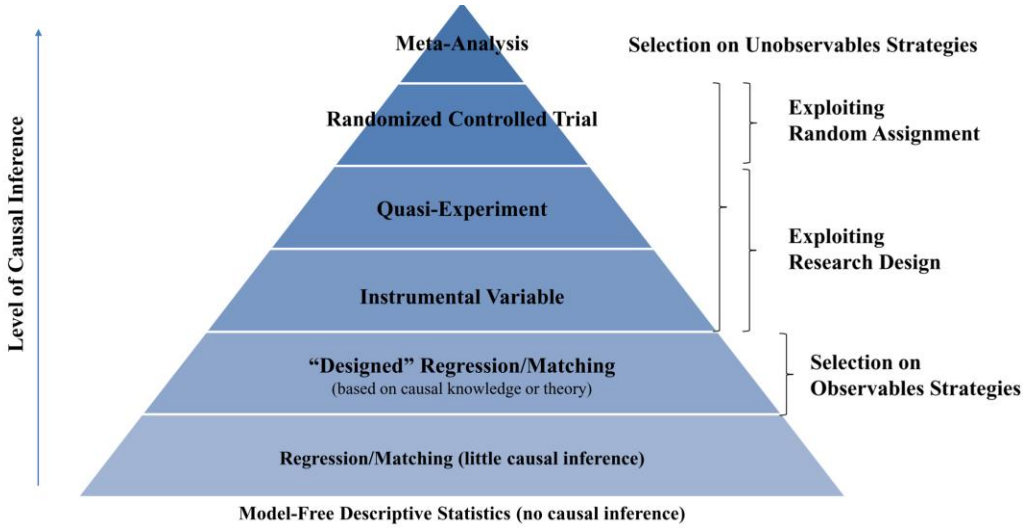
요약

Causal-Lab

1. 연관 관계는 때로는 인과 관계가 될 수 있다.
- 단, 편향이 존재하지 않을 때
- 편향을 제거하는 방법은 추후 배울 내용임

2. 머신러닝도 데이터를 잘 설계하면 인과효과를 추정할 수 있다.
- 7장에서 자세히 배울 예정

3. 인과추론은 식별 과정을 통해 이뤄진다.
- 식별이란? 관측 가능한 데이터에서 인과 추정량을 찾아내는 방법
- 그리고 여기에는 다양한 방법



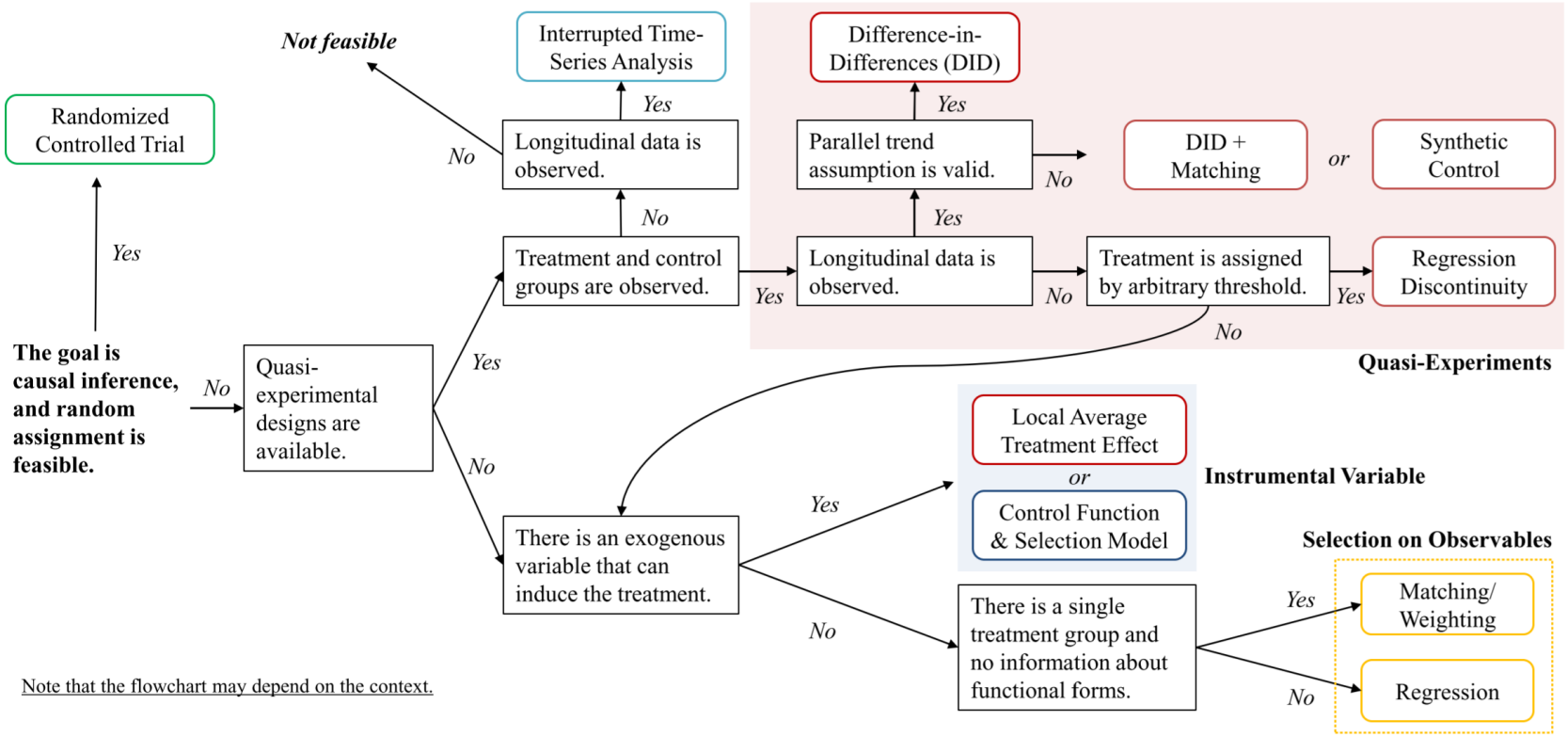
요약

Causal-Lab

1. 연관 관계는 때로는 인과 관계가 될 수 있다.
- 단, 편향이 존재하지 않을 때
- 편향을 제거하는 방법은 추후 배울 내용임

2. 머신러닝도 데이터를 잘 설계하면 인과효과를 추정할 수 있다.
- 7장에서 자세히 배울 예정

3. 인과추론은 식별 과정을 통해 이뤄진다.
- 식별이란? 관측 가능한 데이터에서 인과 추정량을 찾아내는 방법
- 그리고 여기에는 다양한 방법



끝

Causal-Lab

Q&A

발표 정하기

Causal-Lab

주차	날짜	내용	발표자
Week 1	2025-03-03	OT	신진수
Week 2	2025-03-09	Part 1.인과추론의 기초	박이삭
Week 3	2025-03-16	Part 2.선형회귀	미정
Week 4	2025-03-24	Part 3.성향점수	미정
Week 5	2025-03-29	☆가짜연구소 인과추론팀 Meet Up	-
Week 6	2025-04-06	Part 4.이질적 처치효과	미정
Week 7	2025-04-13	Part 5.메타러너	미정
Week 8	2025-04-20	Part 5.이중차분법	미정
Week 8	2025-04-27	☆가짜연구소 Magical Week 휴식	-
Week 8	2025-05-04	Part 5.통제집단 합성법	미정
Week 9	2025-05-11	Part 5.대안적 실험설계	미정
Week 10	2025-05-17	가짜연구소 Pseudo 콘서트	-

강지민
김소희
김시은
김용민
김정현
여홍수