가짜연구소

Causal-Lab

성향 점수

2025. 03. 23

김소희

발표자 소개

Causal-Lab

김소희

데이터 분석가

- 2019 ~ 2020 왓챠(영상 스트리밍)
- 2020 딜리버리랩(식자재 비교, 배송)
- 2021 ~ 티빙(영상 스트리밍)

목차

- 관리자 교육의 효과
- 회귀분석과 보정
- 성향점수
- 디자인 vs. 모델 기반 식별
- 이중 강건 추정
- 연속형 처치에서의 일반화 성향점수

(예시) 관리자 교육의 효과

Causal-Lab

회사에서 좋은 관리자를 육성하기 위해 관리자를 대상으로 한 교육 프로그램을 운영함

교육의 효과를 추정하기 위해 관리자들을 무작위로 교육 프로그램에 참여시키고, 참여한 관리자와 참여하지 않은 관리자의 부하직원들의 참여도를 비교하려 함(무작위 실험 연구)

그러나 교육 프로그램에 배정된(실험군) 일부 관리자가 참석하지 않았고, 반대로 일부 대상이 아닌(대조군) 관리자가 와서 교육을 듣기도 하면서 원래의 의도된 무작위 실험 연구가 관측 연구가 되었다.

Q. 어떻게 하면 교란 요인을 보정해서 실험군과 대조군을 비교해서 교육의 효과를 추정할 수 있을까?

인과 관점에서의 상황 요약

- **처치**: 관리자의 교육 프로그램 참여 여부
- **결과 변수**: 해당 관리자와 일하는 직원의 참여 점수(성과와 관련된 것으로 추정)
- **그 외 공변량**: 관리자 근속기간, 관리자의 보고서 수, 관리자 성별, 회사 내 직군, 부서 직원 수 등등

회귀분석과 보정

Causal-Lab

먼저 4장에서 배운 회귀분석을 사용해서 교란 요인을 보정하여 교육 효과를 추정한다.

OLS로 engagement_score ~ intervention 방식으로 단순 비교 시

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.2347	0.014	-16.619	0.000	-0.262	-0.207
intervention	0.4346	0.019	22.616	0.000	0.397	0.472

결과: 관리자 교육 시 같이 일하는 직원의 참여 점수가 0.4346만큼 증

→ but 이건 처치가 완전히 무작위 배정되었을 때의 이야기이고, 그렇지 않았으므로 이 결과는 편향되었을 것.

이제 데이터에 주어진 다른 공변량도 모두 넣어 모델에 넣어 다시 OLS 모델을 적합해본다.

결과: 관리자 교육 시 같이 일하는 직원의 참여 점수가 0.2677만큼 증가한다.

→ 효과 추정값이 아까보다는 훨씬 작아짐

→ 원인 추정: 이미 직원들의 참여도가 높은 팀의 관리자가 교육 프로그램에 더 많이 참여했을 가능성(긍정 편향)

```
model = smf.ols("""
engagement_score ~ intervention + tenure + last_engagement_score + department_score + n_of_reports + C(gender) + C(role)
""", data=df).fit()
```

```
ATE:   0.26779085766768 64
95% CI:   [0.23357751 0.30200421]
```

성향 점수(Propensity Score)

Causal-Lab

이제 성향점수가 무엇인지 배우고, 앞서 구한 회귀 추정값과 성향점수의 가중치 추정값이 같은지 비교해보자.

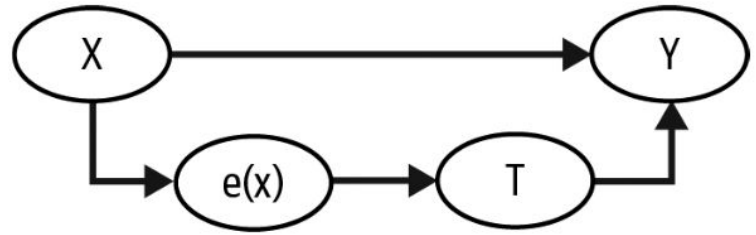
당면한 문제: 교란요인 X로 인해 처치가 무작위 배정되지 않음. 근데 X를 직접 통제할 수가 없었음.

<u>성향점수란</u>? 처치를 받을 조건부 확률 = 공변량 X를 처치 T로 변환하는 일종의 함수

<u>성향점수를 고안한 배경</u>? 교란 요인 X를 직접 통제할 필요 없이 $(Y_1, Y_0) \perp T \mid X$ 를 만족할 수 있겠다!

교란 요인을 통제하는 대신 $E[T \mid X]$ 를 추정하는 균형점수를 통제하면

균형점수 = 처치의 조건부 확률 P(T|X) = 성향점수 e(x)



→ e(x), 즉 처치에 대한 조건부 확률을 알면 굳이 X를 통제하지 않아도 X를 통제할 때와 같은 효과(순수하게 T가 Y에 미친 영향을 아는 것)를 얻을 수 있음.

성향 점수 추정

Causal-Lab

성향 점수를 알면 무작위로 배정되지 않은 처치를 무작위로 배정된 것처럼 만들 수 있는 것을 알았다.
그럼 이제 성향점수를 추정해보자

사실 실제 성향점수는 정확히 알 수 없는 값이고, 추정할 수밖에 없는 값.

로지스틱 회귀를 이용해서(예시의 처치가 binary이기 때문에) 성향점수를 추정해보자.

성향점수 = logit(교육 프로그램 참가 여부(T) ~ 근속연수 + 부서 점수 + 보고 수 + ... + 공변량들(X))

```
# 공변량 X들로 처치 X를 추정하는 logit 모델
ps_model = smf.logit("""
intervention ~ tenure + last_engagement_score + department_score + n_of_reports + C(gender) + C(role)
""", data=df).fit(disp=0)
```

추정된 성향점수

	intervention	engagement_score	propensity_score
0	1	0.277359	0.593887
1	1	-0.449646	0.399909
2	1	0.769703	0.600523
3	1	-0.121763	0.578387
4	1	1.526147	0.617720

성향 점수와 직교화의 유사성

Causal-Lab

성향점수 추정은 선형회귀와 매우 비슷함

OLS도 편향 제거 단계에서 E[T|X] 를 추정한다 = 처치 배정 메커니즘을 모델링 한다.

그럼 선형회귀에서 교란 요인 X를 보정하기 위해 성향점수 $e(X)$ 를 사용할 수도 있지 않을까? => 사용해본다.

OLS(참여 점수(Y) ~ 교육 프로그램 참가 여부(T) + 성향 점수)

→ intervention의 회귀계수: 0.2661

→ 앞서 처치와 교란요인 X를 모두 넣어 적합시킨 회귀계수 0.2677과 유사한 결과

OLS는 선형회귀를 사용해서 T를 모델링했고 성향점수 추정값은 로지스틱 회귀를 사용했다는 차이가 있을 뿐 두 접근 모두 처치를 직교화하는 방법이라는 점에서는 같다.

성향 점수를 활용하여 처치가 무작위 배정된 것처럼 만드는 법

Causal-Lab

1. 성향점수 매칭(PSM)

2. 성향점수를 활용한 역확률 가중치(IPW)

성향 점수 매칭(Propensity Score Matching, PSM)

Causal-Lab

매칭 추정량(matching estimator)를 통해 성향점수를 통제해보자.

추정 방법: 관측 가능한 특징이 비슷한 실험 대상의 짝을 찾아서 실험군과 대조군을 비교한다. 비슷한 실험 대상의 매칭에 사용할 수 있는 알고리즘 종류는 다양한데 여기서는 K=1인 KNN 알고리즘을 사용하여 매칭한다.

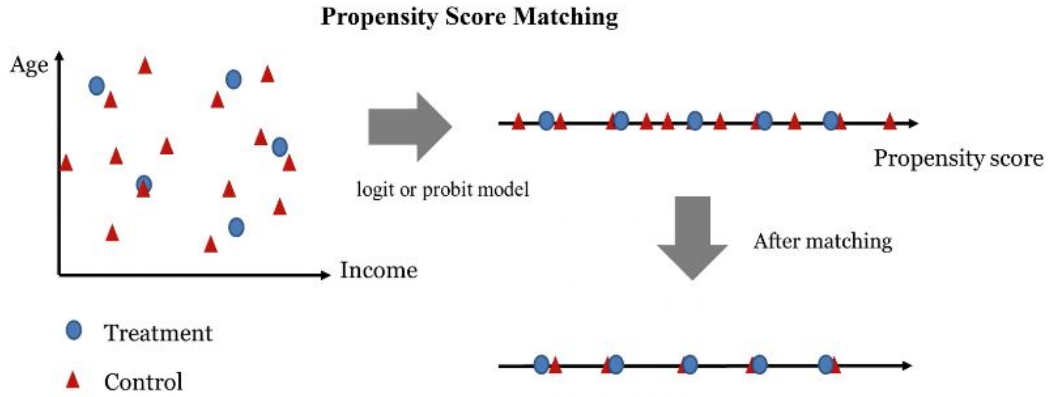
1. 교란 요인이 될 수 있는 공변량들을 처치 여부에 대해 회귀하여 성향점수를 구한다.

2. 성향점수를 유일한 피쳐로 하여 실험군에 대해 KNN 모델을 적합시킨다.

3. 실험군에 대해 K=1인 최근접 이웃(=매칭된 실험 대상)을 구해서 대조군의 Y_1 을 대체한다.

4. 대조군에 대해서도 2, 3의 과정을 반복하여 실험군의 Y_0 를 대체한다.

5. 각 실험 대상에 짝이 지어졌다면, 매칭된 대조군과 실험군의 평균을 비교하여 ATE를 추정한다.



$$\hat{ATE} = \frac{1}{N}\Sigma(Y_i - Y_{jm}(i))T_i + (Y_{jm}(i) - Y_i)(1 - T_i)$$

그림 출처: Korea Summer Workshop on Causal Inference 2022 부트캠프 '매칭과 역확률 가중치' 발표 자료

성향 점수 매칭(Propensity Score Matching, PSM)

Causal-Lab

매칭 추정량(matching estimator)를 통해 성향점수를 통제해보자.

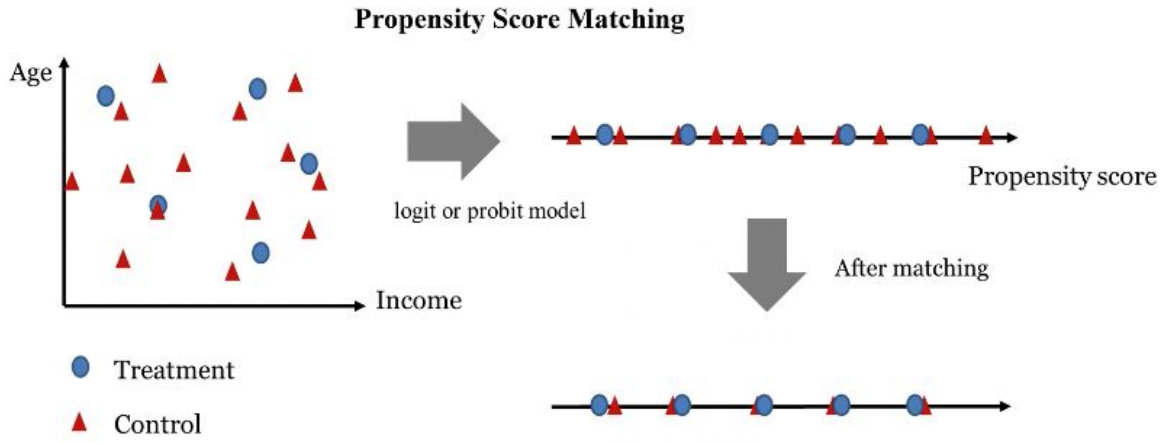


그림 출처: Korea Summer Workshop on Causal Inference 2022 부트캠프 '매칭과 역확률 가중치' 발표 자료

$$\hat{ATE} = \frac{1}{N}\Sigma(Y_i - Y_{jm}(i))T_i + (Y_{jm}(i) - Y_i)(1 - T_i)$$

c.f. 정확 매칭(exact matching): 성향점수로만 매칭을 해야만 할까? → NO

어떤 함수를 가정하거나 propensity score를 구하지 않고 원래의 특성 X를 그대로 사용해서 매칭하는 방법. 그러나 X의 차원이 클수록 데이터가 희소해지고 매칭의 정확성이 더 떨어지는 문제가 있어, 편향 보정식을 적용해야 한다.

c.f. 정확 매칭(Exact Matching)

Causal-Lab

성향점수로만 매칭을 해야만 할까? → NO

정확 매칭: 어떤 함수를 가정하거나 propensity score를 구하지 않고 원래의 특성 X를 그대로 사용해서 매칭하는 방법

그러나 X의 차원이 클수록 데이터가 희소해지고 매칭의 정확성이 더 떨어지는 문제가 있다.

따라서 exact matching을 할 때에는 다음과 같이 실험군의 편향과 대조군의 편향을 제거해주어야 한다.

$$\widehat{ATE} = \frac{1}{N}\sum\left\{\left(Y_i - Y_{jm}(i) - \left(\hat{\mu}_0(X_i) - \mu_0(X_{jm})\right)\right)T_i + \left(Y_{jm}(i) + Y_i - \left(\hat{\mu}_1(X_{jm}) - \mu_1(X_i)\right)\right)(1-T_i)\right\}$$

Coarsened Exact Matching(CEM)

공변량을 그룹화(=coarsening) 함으로써 성향점수 매칭과 정확 매칭의 중간적 성격을 띄도록 만드는 방식

성향 점수 매칭의 단점

Causal-Lab

저자가 PSM을 꺼려하는 이유

1. 편향될 가능성이 있다.

2. 분산을 추정하기 어렵다.

3. 데이터 과학 활용 경험에서 KNN은 X가 고차원인 경우 효율이 크게 떨어진다 → 차원의 저주
(이건 exact matching에만 해당하는 문제)

Gary King & Richard Nielsen. 2019. Why Propensity Scores Should Not Be Used for Matching. Political Analysis, 27, 4

- PSM에서는 성향점수 추정 시 포함되지 않은 교란변수나 상호작용 효과로 인해 그룹 간 차이가 생길 수 있다.
- PSM은 원래의 목적과 달리 현실적으로 연구자가 하나의 모델을 선택하기 위해서 많은 모델을 시도하게 만든다.

성향 점수를 활용하여 처치가 무작위 배정된 것처럼 만드는 법

Causal-Lab

1. 성향점수 매칭(PSM)
2. **성향점수를 활용한 역확률 가중치(IPW)**

역확률 가중치

Causal-Lab

표본에 처치받을 확률의 역확률을 가중치로 곱해서, 모든 실험 대상이 처치 t를 받았을 경우와 비슷한 유사 모집단(pseudo-population) 생성

처치군에 대한 역확률: $\frac{1}{P(T_i=1|X_i)}$

대조군에 대한 역확률: $\frac{1}{P(T_i=0|X_i)}$

(예시) 모든 관리자가 교육을 받았을 때의 평균 참여도(Y_i) 의 기댓값을 구하고 싶을 때

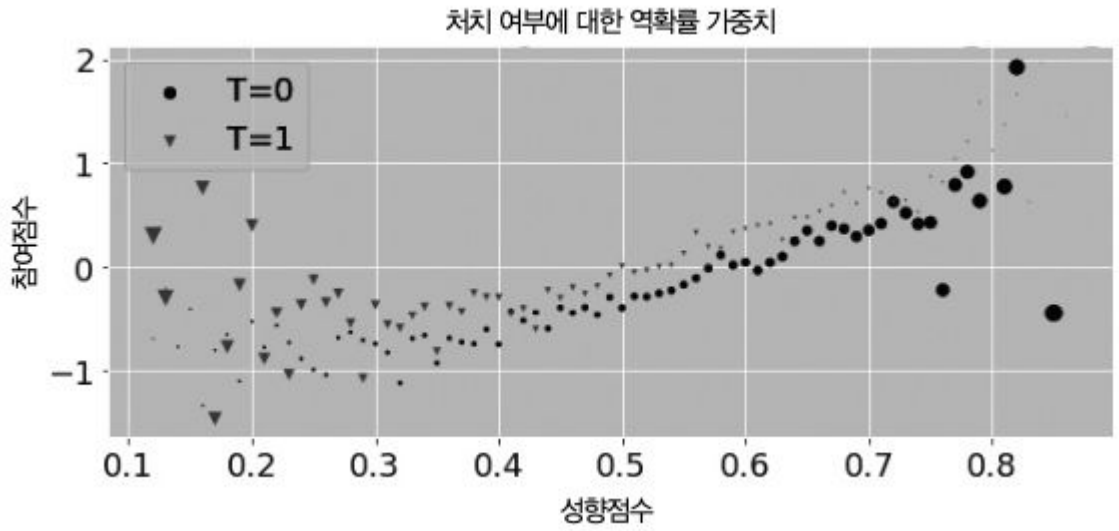
- 실험군(처치=교육받음)을 처치 받을 확률의 역수를 곱해서 조정한다
  - 처치받을 확률이 낮았다 → 분모가 작아지므로 역확률은 커짐
  - 처치받을 확률이 컸다 → 분모가 커지므로 역확률은 작아짐
- 즉 처치 받을 확률이 매우 낮은데도 처치를 받은(실질적으로 드물게 일어난 처치 사례) 대상에게는 높은 가중치를 부여하는 것. 반대로 처치 받을 확률이 매우 큰데 처치를 받지 않은 사례에 대해서도 큰 가중치를 부여.

실험군 중에는 대조군과 유사한 특성을 가진 대상에, 대조군 중에는 실험군과 유사한 특성을 가진 대상에 높은 가중치를 부여함으로써 두 그룹 간 공변량 분포를 유사하게 만들고, 처치 효과를 추정함에 있어 공변량의 영향을 적게 만드는 방식

역확률 가중치

Causal-Lab

관리자 교육 훈련 데이터의 IPW 과정 그래프



- 성향점수가 낮은 구간의 크기가 큰(가중치가 큰) 역삼각형: 교육받지 않은 것처럼 보이는 교육받은 관리자 $Y_0|T=1$ 에 대한 정보 제공

- 성향점수가 높은 구간의 크기가 큰(가중치가 큰) 동그라미: 교육받은 것처럼 보이는 교육받지 않은 관리자 $Y_1|T=0$ 에 대한 정보 제공

역확률 가중치로 ATE 계산

Causal-Lab

ATE = 실험군에 대해 역확률을 곱해 만든 평균 잠재적 효과 - 대조군에 대해 역확률을 곱해 만든 평균 잠재적 효과

$$ATE = E\left[\frac{\mathbf{1}(T=1)Y}{P(T=1 \mid X)}\right] - E\left[\frac{\mathbf{1}(T=0)Y}{P(T=0 \mid X)}\right]$$

역확률 가중치의 표준오차 계산 - 부트스트랩 방법

Causal-Lab

IPW 추정값의 신뢰구간을 얻는 가장 간단한 방법은 부트스트랩bootstrap 방식

데이터를 반복적으로 복원추출해서 여러 IPW 추정값을 구한 후, 추정값의 2.5번째와 97.5번째 백분위수를 계산해서 95% 신뢰구간을 얻을 수 있다.

주의점

- 큰 가중치가 적용되면 성향점수 추정량의 분산도 크게 증가한다.
  - 최종 추정값에 큰 영향을 주는 소수의 대상들이 분산 증가의 원인
- 성향점수가 높은 부분에 대조군의 실험 대상이 적건, 성향점수가 낮은 부분에 실험군의 대상이 적은 경우 반사실적 결과인 $Y|T=1$ 와 $Y|T=0$ 를 추정할 대상이 적어지므로 결과에 잡음이 많아지는 것

안정된 성향점수 가중치

Causal-Lab

머신러닝 관점에서 IPW를 중요도 샘플링(importance sampling)의 응용으로 볼 수도 있다.

- 중요도 샘플링에는 원하는 타겟 분포 p(x)에서 직접 샘플링이 어려운 경우 원본 분포 q(x)의 데이터에서 샘플링을 수행한 후 q(x)의 데이터에 p(x)/q(x)를 곱해 재조정하는 방식을 사용
- IPW에서 실험군에 1/P(T=1|X)의 가중치를 줘서 P(T=1|X) 분포에서 나온 데이터를 사용하되 이 데이터를 원하는 P(T=1) = 1의 분포로 재구성하는 방식이라는 점에서 유사

그런데 처치 확률이 매우 작으면 P(T|X) 값이 매우 작아지고 → 그럼 가중치가 너무 커져서 계산상 문제가 발생할 수 있음

원래의 가중치에 P(T=t)를 곱해줘서 안정화된 가중치(stabilized weight)를 만들어주면 분모와 분자가 균형을 이뤄서 계산할 때 좀 더 안정적. $W_i = \frac{P(T=1)}{P(T=1|X_i)}$

- 실험군에 대한 안정화된 가중치:

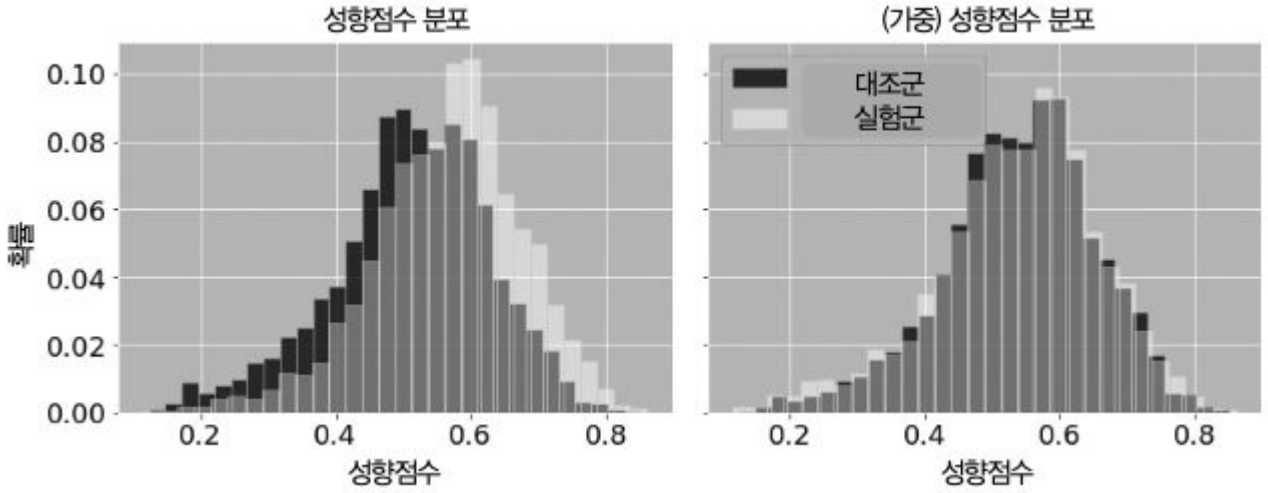
$$W_i = \frac{P(T=0)}{1-P(T=1|X_i)}$$

- 대조군에 대한 안정화된 가중치:

유사 모집단(pseudo-population)

$P(T|X)$의 관점에서 편향의 의미

- 처치가 10% 확률로 무작위 배정되었다 → 처치는 X에 독립 $P(T)$ $P(T|X)$ = 10%
  - 처치와 X가 독립이면 X에서 오는 교란편향이 없고 보정할 필요도 없음
  - 반면 처치와 X가 독립이 아니라면, 일부 실험 대상이 다른 대상에 비해 처치 받을 확률이 높다.



- 왼쪽 그림: 처치가 무작위 배정 되지 않고 X로 인한 교향 편향이 있으면 처치받을 확률이 동일하지 않으므로 처치를 받은 $\hat{e}(x)$ 사람들의 가 더 높게 분포한다.
- 오른쪽 그림: 가중 성향점수 분포에서는 $\hat{e}(x)$ 가 낮은 실험군에 대해 높은 가중치를, 대조군은 낮은 가중치를 주는 보정을 함으로써 두 분포를 비슷하게 만든다 = 실험군과 대조군이 처치 받거나 처치 받거나 받지 않을 확률이 같아진다 = 처치 배정이 마치 무작위인 것처럼 보이게 만든다.

IPW로 선택편향 보정하기

Causal-Lab

IPW는 선택 문제(selection issue)를 보정하는 데에도 사용 가능하다.

(예시) 앱에 대한 고객의 만족도를 알아보기 위해 1~5 척도로 제품을 평가하는 설문 진행

- 일부 고객이 설문에 응답하지 않음 → 설문 응답자와 미응답자 간 차이로 인해 설문 결과 편향 가능
- 예를 들어 앱에 불만족한 고객들이 대부분 설문에 응답하지 않았다면 설문 결과는 만족하는 쪽으로 편향됨
- 이를 보정하기 위해 고객의 공변량(나이, 소득, 앱 사용량 등)을 이용해서 응답률 R과 P(R=1|X)를 추정할 수 있다. 그러고 나서 응답자에게 $1/\hat{P}(R=1)$ 만큼의 가중치를 부여함
  - 응답자 중 $\hat{P}(R=1)$ 이 낮은, 즉 미응답자와 유사한 응답자에게 높은 가중치를 부여해서 마치 모두가 설문에 응답한 것처럼 보이는 유사 모집단을 생성한다.
  - 즉 응답자 중에서도 미응답자와 유사한 특성을 가진 응답자에게 높은 가중치를 부여해서 (미응답자의 특성을 포함해) 앱을 쓰는 모든 고객의 특성을 반영한 것 같은 유사 모집단을 생성하는 것

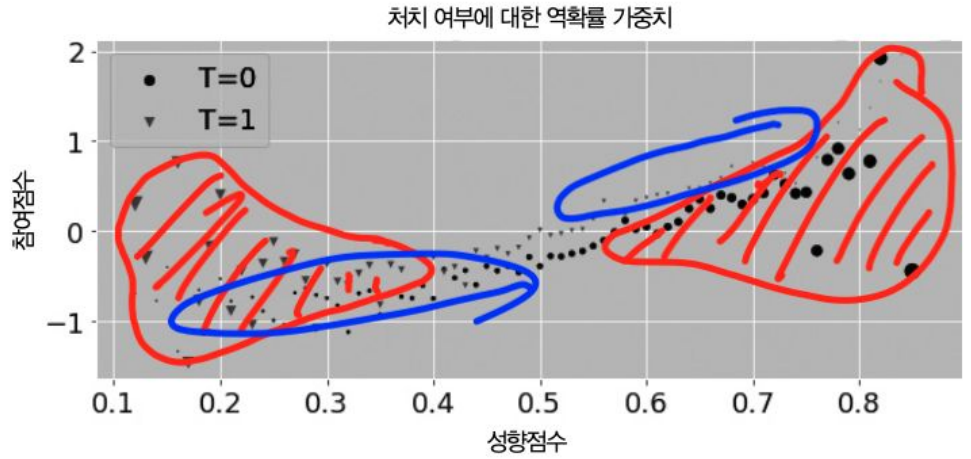
성향점수의 한계: 편향-분산 트레이드오프

Causal-Lab

4.9.1에서 본 '잡음 유발 통제변수'처럼, 성향 점수를 너무 잘 예측하는 모델은 분산이 너무 커져 부정확한 효과 추정값을 생성할 수도 있다.

만약 T를 굉장히 잘 예측하는 모델이 있어서 아래와 같이 나타난다고 해보자.

- 모든 실험군에 대해서 $\hat{e}(x)$ 가 매우 높다 = 실험군이 처치 받은 대상임을 정확하게 예측
- 모든 실험군에 대해서 $\hat{e}(x)$ 가 매우 낮다 = 대조군이 처치 받지 않은 대상임을 정확하게 예측



→ $Y_1|T=0$를 추정할 수 있는 $e(x)$ 이 낮은 실험군이 없고 $Y_0|T=1$ 를 추정할 수 있는 $\hat{e}(x)$ 이 높은 대조군이 없어짐

→ 얼마 없는 성향점수가 낮게 예측된 실험군과 성향점수가 높게 예측된 대조군에 더 큰 가중치가 부여되면서 분산이 증가한다.

디자인 vs. 모델 기반 식별

Causal-Lab

- 모델 기반 식별
  - 처치 및 추가 공변량을 조건부로 설정하여 잠재적 결과에 대한 모델 형태로 가정
  - 목표: 추정에 필요한, 누락된 잠재적 결과를 대체하자.
  - (예시) 통제집단합성법, 회귀분석
- 디자인 기반 식별
  - 처치 배정 메커니즘에 대한 가정
  - 이 책에서는 통계적 모델링이 아닌 연구 디자인에 기반한 인과추론도 포함시킴
  - (예시) IPW, 회귀분석

→ 처치 배정 메커니즘과 잠재적 결과 모델 중 어떤 가정이 더 편한지에 대한 답에 따라 선택할 수 있다.

이중 강건 추정(doubly robust, DR)

Causal-Lab

모델 기반과 디자인 기반 식별을 결합해서 적어도 둘 중 하나가 정확하면 처치효과를 정확히 추정할 수 있게 하는 방식

- 성향 점수 모델: $E[T|X]$ → P(T|X)의 성향점수
- 결과 모델: $\hat{m}(X)$ 를 추정하는 모델 ← 선형회귀 사용 $\hat{e}(x)$

아래의 반사실 Y_t에 대한 이중 강건 추정량 식 (는 결과모델, 는 성향점수)

$$\hat{\mu}_t^{DR}(\hat{m},\hat{e}) = \frac{1}{N}\sum \hat{m}(X) + \frac{1}{N}\sum\left[\frac{T}{\hat{e}(x)}(Y-\hat{m}(X))\right] \qquad \hat{\mu}_t^{DR}(\hat{m},\hat{e}) = \frac{1}{N}\sum\frac{TY}{\hat{e}(X)} - \frac{1}{N}\sum\left[\frac{T-\hat{e}(X)}{\hat{e}(X)}\hat{m}(X)\right]$$

- 성향점수 모델이 잘못되었으나 결과 모델이 정확한 경우 → 두번째 항에서 $E[Y-\hat{m}(X)]=0$ 이므로 첫째 항의 결과 모델만 남게 됨
- 결과 모델이 잘못되었으나 성향점수 모델이 정확한 경우 → 전개한 식에서 두번째 항의 $E[\hat{e}(x)]=0$ 으로 수렴하여 IPW 추정량만 남게 됨

→ 둘 중 하나의 모델만 맞아도 평균 반사실적 결과인 Y_t 가 수렴한다.

→ E[Y_0]와 E[Y_1] 각각에 두 추정량을 사용하고 그 차이를 계산해서 평균 처치 효과 추정

$$ATE = \hat{\mu}_1^{DR}(m,e) - \mu_0^{DR}(m,e)$$

연속형 처치에서의 일반화 성향점수

Causal-Lab

4장에서는 연속형 처치에 대해 처치 반응의 함수 형태를 가정해서 문제를 해결했지만, 성향점수 가중치에서는 모수적 반응 함수 같은 것은 없고, 잠재적 결과는 재조정하고 평균을 구하는 비모수적 방식으로 추정된다.

- T가 연속형일 때 잠재적 결과 Y_t 는 무한히 많이 존재한다.
- 연속형 변수의 확률은 항상 0(확률 밀도 함수에서 한 점의 면적은 0이기 때문에) 이므로 $P(T=t|X)$ 를 추정하는 것도 불가능

이런 문제를 해결하는 방법은

- 연속형 처치를 이산화 하기
- 일반화 성향 점수(generalized propensity score, GPS)를 사용하기

정리

Causal-Lab

- 역확률 가중치를 이용하여 교란변수로 인해 생긴 데이터의 편향을 보정할 수 있다.
- 회귀분석 및 직교화와 마찬가지로 처치 배정 메커니즘을 모델링하는 디자인 기반 식별 방법의 일종이다.
- 직교화가 처치를 잔차화 했다면, IPW는 처치의 차원은 유지하되 각 데이터를 처치 성향점수의 역수로 재조정함으로써 처치가 공변량 X에 독립인 분포에서 추출된 것처럼 보이게 만든다.
- 선형회귀와 IPW 중 어떤 것을 언제 선택해야 할까?

  - 처치가 이산형 → IPW를 사용하면 좋다.

  - 처치가 연속형 → 특정 처치 주변에 데이터가 적을 수 있기 때문에 회귀 모델링이 더 유리하다.

  - IPW를 쓸 때에는 이중 강건으로 결과 모델과 함께 사용하면 좋다.

감사합니다