



2. 인과 추론에 대한 몇 가지 세부 사항 + Python의 구체적인 예제

🕒 생성일	@2022년 2월 3일 오후 2:25
☰ 태그	

[Where to start](#)

[What is Causal Inference?](#)

[3 Gifts of Causal Inference](#)

[Gift 1: do-operator](#)

[Gift 2: Deconfounding Confounding](#)

[Gift 3: Estimating Causal Effects](#)

[Conclusion](#)

[참고](#)

Where to start

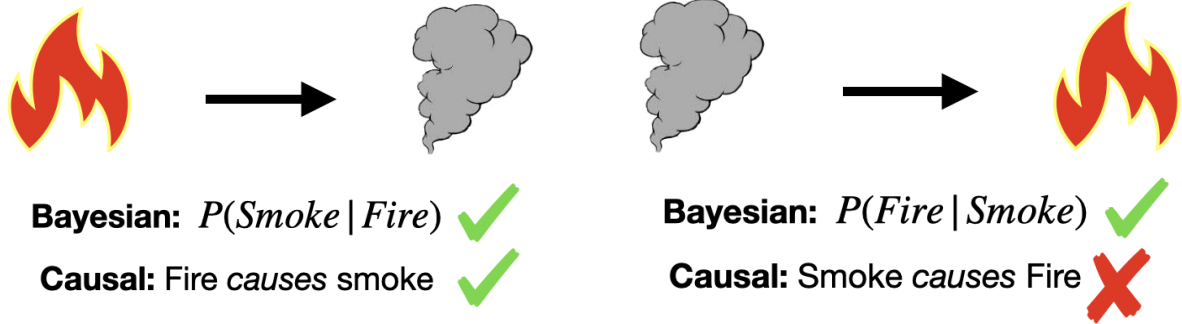
지난 게시물에서 SCM(구조적 인과 모델)을 통해 인과 관계를 수학적으로 표현하는 방법에 대해 논의했습니다. SCM은 두 부분으로 구성됩니다. 인과 관계를 시각화하는 그래프와 연결된 부분들의 세부 사항을 표현하는 방정식, 등식 (equations) 입니다.

요약하자면, 그래프는 꼭짓점(노드)과 모서리(링크)로 구성됩니다. 여기서는 **그래프와 네트워크**라는 용어를 같은 의미로 사용하겠습니다. SCM은 DAG(Directed Acyclic Graph)라고 하는 특별한 종류의 그래프를 사용합니다. 이 그래프에는 모든 모서리가 한 방향을 향하고, 사이클이 없습니다. DAG는 인과 추론을 위한 출발점입니다.

Bayesian vs Causal Networks

먼저 간단하게, 베이지안과 케주얼 네트워크의 차이점을 설명 드리겠습니다.

표면적으로 베이지안 네트워크와 인과 네트워크는 완전히 동일합니다. 그러나 차이점은 해석에 있습니다.



여기에 2개의 노드(불 아이콘과 연기 아이콘)와 1개의 엣지 (불에서 연기를 가리키는 화살표)가 있는 네트워크가 있습니다. 이 네트워크는 베이저안 네트워크일 수도 있고 인과 네트워크일 수도 있습니다.

그러나 주요 차이점은 이 네트워크를 해석할 때 생깁니다. 베이저안 네트워크의 경우 노드를 변수로, 화살표를 조건부 확률, 즉 화재에 대한 정보가 주어졌을때의 연기 확률로 봅니다. 이것을 인과 네트워크로 해석할 때, 노드는 여전히 변수로 간주되지만 화살표는 인과 관계를 나타냅니다.

이 경우 두 해석이 모두 유효합니다. 그러나 엣지 (화살표) 방향을 뒤집으면 연기가 화재를 일으키지 않기 때문에 인과 네트워크 해석은 옳지 않게 됩니다.

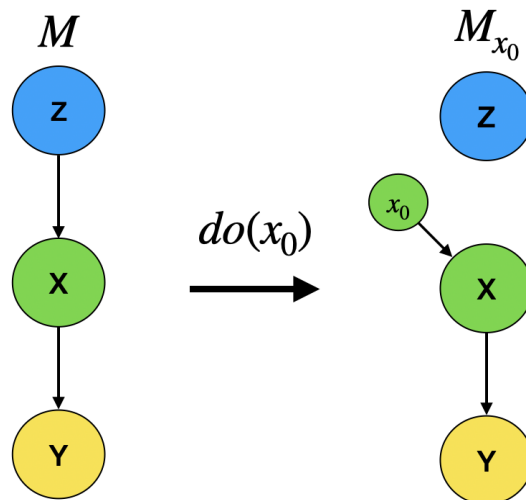
What is Causal Inference?

인과 추론은 단순한 통계적 질문이 아닌 **인과 관계 인지 아닌지**에 대한 질문에 답하는 것을 목표로 합니다. 인과 추론의 적용은 무수히 많습니다. 아래 질문에 답하는 것은 인과관계 추론의 범주에 속합니다

- 치료가 복용한 사람들에게 직접적인 도움이 되었습니까?
- 이번 달이나 휴일에 매출 증가로 이어진 마케팅 캠페인이 있습니까?
- 임금 인상이 생산성에 얼마나 큰 영향을 미칠까요?

이는 보다 전통적인 접근 방식 (예: 선형 회귀 또는 표준 기계 학습)을 사용하여 쉽게 대답할 수 없는 중요하고 실용적인 질문입니다. 나는 내가 **인과 추론의 3가지 선물**이라고 부르는 것을 통해 인과 추론이 이러한 질문에 답하는 데 어떻게 도움이 될 수 있는지 설명하고자 합니다.

3 Gifts of Causal Inference



Gift 1: do-operator

지난 포스트에서 저는 해결책의 관점에서 인과관계를 정의했습니다. 몇 가지 전문적인 내용을 생략하고 X에 대한 개입이 Y를 변경하는 경우 X가 Y를 유발하는 반면, Y에 대한 개입이 반드시 X의 변경을 초래하는 것은 아닙니다.

개입은 실제 세계에서 이해하기 쉽습니다. 그러나 이것이 어떻게 인과 관계의 수학적 표현에 들어맞습니까?

do-operator는 물리적 개입의 수학적 표현입니다. $Z \rightarrow X \rightarrow Y$ 모델로 시작하면 들어오는 모든 화살표를 X로 삭제하고 X를 일부 값 x_0 으로 수동으로 설정하여 X에 대한 개입을 시뮬레이션할 수 있습니다. (위 그림 참조)

do-operator의 힘은 인과 관계의 세부 사항을 알고 있는 경우 실험을 시뮬레이션할 수 있다는 것입니다. 예를 들어 마케팅 예산을 늘리면 매출이 늘어날까? 와 같은 질문을 던질 수 있습니다.

마케팅 지출과 판매를 포함하는 인과 모델로 만들고, 마케팅 지출을 늘리면 어떻게 될지 시뮬레이션 한뒤 판매 변화에 가치가 있는지 평가할 수 있습니다. 즉, 마케팅이 매출에 미치는 인과적 효과를 평가할 수 있습니다.

Pearl과 동료들의 주요 공헌은 **Do-calculus**의 규칙입니다. 이것은 do-operator를 사용하는 방법을 설명하는 완전한 규칙 (**set of rules**)입니다. 특히 do-calculus는 중재하는 **interventional** 분포(do-operator가 있는 확률)를 관찰 **observational** 분포(do-operator가 없는 확률)로 변환할 수 있습니다. 이는 아래 그림의 규칙 2와 3에서 확인할 수 있습니다.

Rules of Do-Calculus:

1. Insertion/deletion of observations

$$P(Y | do(X), Z, W) = P(Y | do(X), Z)$$

If W is irrelevant to Y

2. Action/observation exchange

$$P(Y | do(X), Z) = P(Y | X, Z)$$

If Z blocks all back-door paths from X to Y

3. Insertion/deletion of actions

$$P(Y | do(X)) = P(Y)$$

If there is no causal path from X to Y

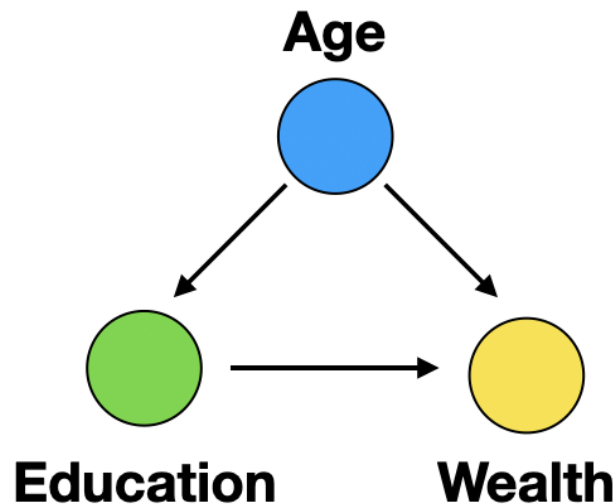
$P(Y|X)$ 는 우리 모두에게 친숙한 조건부 확률, 즉 X에 대한 **observation**이 주어졌을 때 Y의 확률입니다.

반면 $P(Y|do(X))$ 는 do(x)라는 X의 **intervention**이 주어졌을 때 Y의 확률입니다.

do-operator는 인과 추론 도구 상자의 핵심 도구입니다. 사실, 다음 2개의 선물은 do-operator에 의존합니다.

Gift 2: Deconfounding Confounding

혼동 **Confounding** 은 통계에서 흔히 볼 수 있는 개념입니다. 이름을 부르지는 않았지만 심슨의 역설을 통해 이전 게시물에 등장했습니다. 아래 그림은 혼동 **Confounding** 의 간단한 예를 보여줍니다.



이 예시에서 나이 **Age** 는 교육 **education** 과 자산 **Wealth** 의 혼동 요인입니다. (영어를 모르진 않겠지만 헛갈리지 않게 하기 위해 넣었습니다.) 다시 말해, 교육이 부에 미치는 영향을 평가하기 위해서는 연령에 맞게끔 조정할 필요성이 있습니다. 연령을 조정 (또는 조건을 두는 것은)하는 것은 연령, 교육 및 자산 데이터를 볼 때, 연령 그룹이 아닌 **연령 그룹 내의 데이터 포인트를 비교한다는 것**을 의미합니다.

연령이 조정되지 않으면 교육이 부의 진정한 원인인지 아니면 부의 상관 관계인지 명확하지 않을 것입니다. 다시 말해, 교육이 부에 직접적인 영향을 미치는지 아니면 공통된 원인이 있는지 알 수 없습니다.

간단한 예를 들어 보겠습니다. DAG를 볼 때 혼동(confounding)을 일으키는 것입니다. 3개의 변수에 대해 교란자 (confounder)는 2개의 다른 변수를 가리키는 변수입니다.

그러나 더 복잡한 문제는 어떻게 처리합니까? 여기서 **do-operator**는 명확성을 제공합니다. Pearl은 명확한 방식으로 혼란을 정의하기 위해 do-operator를 사용합니다. 그는 혼동이란 $P(Y|X)$ 가 $P(Y|do(X))$ 과는 다르게 변화되는 모든 것이라고 말합니다.

Gift 3: Estimating Causal Effects

이 마지막 선물은 인과 추론의 주요 포인트입니다. 삶에서 우리는 스스로에게 *why* 뿐만 아니라 *how much?* 또한 물어봅니다. 인과관계를 추정하는 것은 이 두 번째 질문 (*how much?*)에 답하는 것으로 귀결됩니다.

대학원을 생각해봅시다. 대학원 학위를 가진 사람들이 대학원 학위가 없는 사람들보다 (대부분) 더 많은 돈을 번다는 것은 한 가지 사실이지만, 그로부터 나올 수 있는 자연스러운 질문은 받는 소득 중 얼마나 많은 돈이 학위로 부터 나오느냐 입니다. 즉, 대학원 학위가 소득에 미치는 효과는 무엇인가? 입니다.

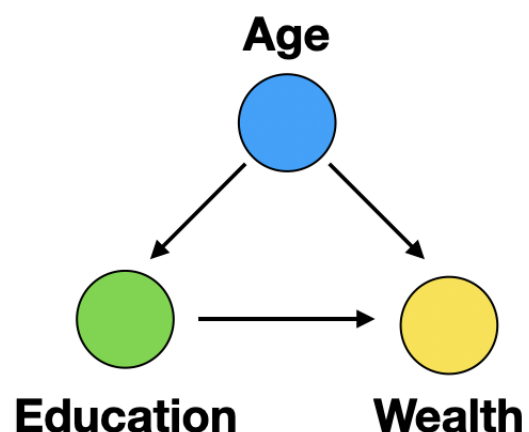
이 질문에 대한 답은 파이썬을 사용하여 인과 추론을 수행하고 구체적인 예를 살펴해보도록 하겠습니다.

Example: Estimating Treatment Effect of Grad School on Income

이 예에서는 인과 추론을 위해 Microsoft DoWhy 라이브러리를 사용합니다. 여기에서 목표는 연간 \$50,000 이상의 수입에 대한 대학원 학위의 인과 효과를 추정하는 것입니다. 데이터는 UCI 머신 러닝 저장소에서 얻습니다. 예제 코드와 데이터는 GitHub 리포지토리에서 찾을 수 있습니다.

모든 인과 추론의 출발점이 인과 모델임을 강조하는 것이 중요합니다. 여기서 우리는 소득이 나이와 교육이라는 두 가지 원인만 가지고 있다고 가정합니다.

여기서 나이도 교육의 원인입니다. 물론 분명히 이 단순한 모델에는 다른 중요한 요소가 누락 되었을 수도 있습니다. 우리는 인과 관계 발견에 관한 다음 포스트에서 **대안 모델을 조사할** 것입니다. 그러나 지금은 이 단순화된 경우에 초점을 맞출 것입니다.



```
# Import libraries
import pickle
import matplotlib.pyplot as plt

import econml
import dowhy
```

```
from dowhy import CausalModel
# Load Data
df = pickle.load( open( "df_causal_inference.p", "rb" ) )
```

라이브러리를 불러옵니다. 라이브러리가 없으면 리포지토리에서 requirements.txt를 확인하세요.

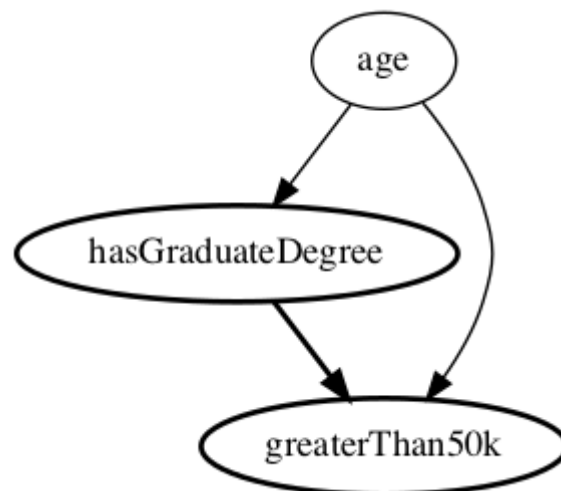
첫 번째 단계는 인과 관계 모델, 즉 DAG를 정의하는 것입니다. DoWhy를 사용하면 모델을 쉽게 만들고 볼 수 있습니다.

```
# Define causal model

model=CausalModel(data = df,
                  treatment= "hasGraduateDegree",
                  outcome= "greaterThan50k",
                  common_causes="age",
                  )

# View model

model.view_model()
from IPython.display import Image, display
display(Image(filename="causal_model.png"))
```



다음으로 추측이 필요합니다.

이것은 기본적으로 우리가 원하는 인과 관계를 제공하는 레시피입니다. 즉, 교육이 소득에 미치는 영향을 계산하는 방법을 알려줍니다.

```
# Generate estimand
identified_estimand= model.identify_effect(proceed_when_unidentifiable=True)
print(identified_estimand)
```

```

Estimand type: nonparametric-ate

### Estimand : 1
Estimand name: backdoor
Estimand expression:
    d
    -----(Expectation(greaterThan50k|age))
d[hasGraduateDegree]
Estimand assumption 1, Unconfoundedness: If U→{hasGraduateDegree} and U→greaterThan50k then P(greaterThan50k|hasGraduateDegree,age,U) = P(greaterThan50k|hasGraduateDegree,age)

### Estimand : 2
Estimand name: iv
No such variable found!

### Estimand : 3
Estimand name: frontdoor
No such variable found!

```

Output of estimand generation

마지막으로 추정치를 기반으로 인과 효과를 계산합니다. 여기에서 EconML 라이브러리의 메타 학습기를 사용하여 별개의 타겟들에 대한 조건부 평균 처리 효과를 추정합니다.

```

# Compute causal effect using metalearner
identified_estimand_experiment = model.identify_effect(
    proceed_when_unidentifiable=True
)

from sklearn.ensemble import RandomForestRegressor

metalearner_estimate = model.estimate_effect(identified_estimand_experiment,
    method_name="backdoor.econml.metalearners.TLearner"
    ,confidence_intervals=False
    ,method_params={
        "init_params":{"models": RandomForestRegressor()},
        "fit_params":{}
    })

print(metalearner_estimate)

*** Causal Estimate ***

## Identified estimand
Estimand type: nonparametric-ate

### Estimand : 1
Estimand name: backdoor
Estimand expression:
    d
    -----(Expectation(greaterThan50k|age))
d[hasGraduateDegree]
Estimand assumption 1, Unconfoundedness: If U→{hasGraduateDegree} and U→greaterThan50k then P(greaterThan50k|hasGraduateDegree,age,U) = P(greaterThan50k|hasGraduateDegree,age)

## Realized estimand
b: greaterThan50k-hasGraduateDegree+age
Target units: ate

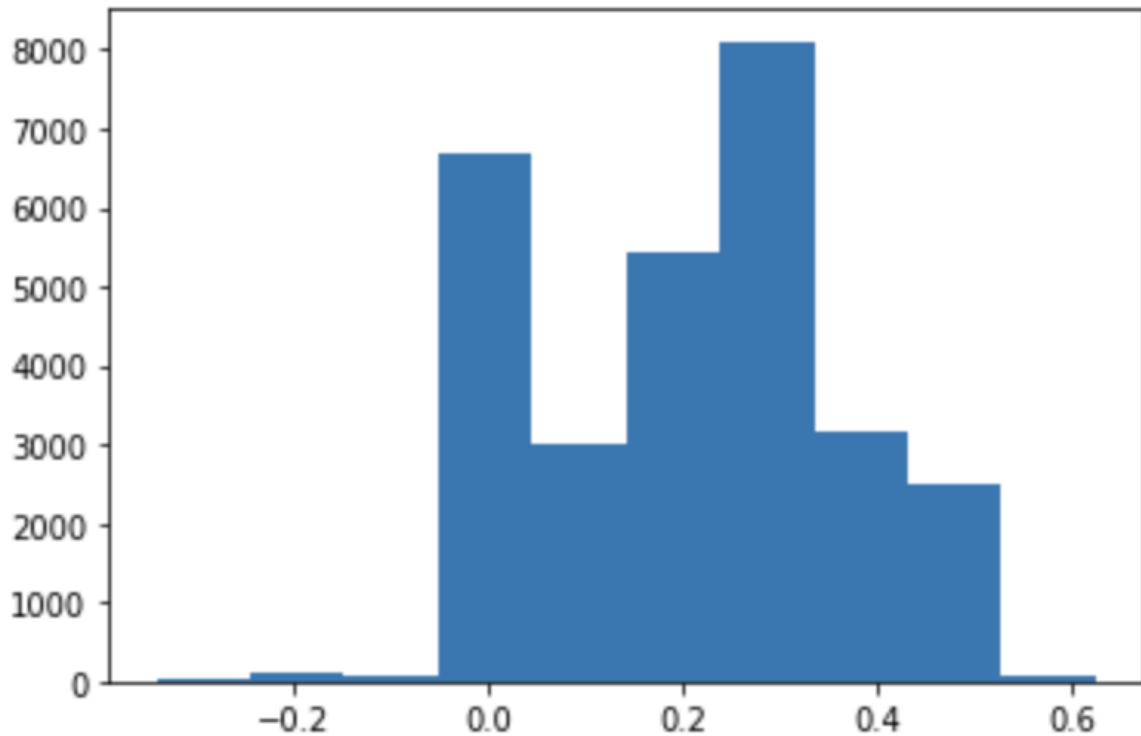
## Estimate
Mean value: 0.20340255646236685
Effect estimates: [ 0.31961958  0.20493831  0.35577517 ...  0.15907199 -0.01266913
 0.19505072]

```

평균 인과 효과는 약 0.20입니다. 이것은 대학원 학위를 취득하면 연간 \$50,000 이상을 벌 확률이 20% 증가한다는 의미로 해석될 수 있습니다. 이것이 평균 효과라는 점에 유의하여 평균이 대표성인지 여부를 평가하기 위해 값의 전체 분포를 고려하는 것이 중요합니다.


```
# Print histogram of causal effects

plt.hist(metalearner_estimate.cate_estimates)
```



위의 그림에서 우리는 표본에 걸친 인과관계의 분포를 볼 수 있습니다. 분명히 분포는 가우스 분포(정규 분포)가 아닙니다. 이는 평균이 전체 분포를 대표하지 않는다는 것을 알려줍니다.

인과 관계에 기반한 코호트에 대한 추가 분석은 **"누가"가 대학원 학위에서 가장 많은 혜택을 받는지**에 대한 실행 가능한 정보를 찾는 데에 도움이 될 수 있습니다.

Conclusion


인과 추론은 기존 접근 방식으로는 해결할 수 없는 자연스러운 질문에 답하기 위한 강력한 도구입니다. 여기에서 인과적 추론에서 몇 가지 큰 아이디어를 스케치하고 코드를 사용하여 구체적인 예를 살펴보았습니다.

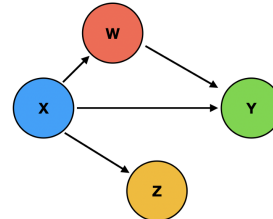
앞에서 언급했듯이 모든 인과 추론의 출발점은 인과 모델입니다. 그러나 일반적으로 우리는 좋은 인과 관계 모델을 가지고 있지 않습니다. 여기에서 인과관계 발견이 도움이 될 수 있으며, 이는 다음 포스트의 주제입니다.

참고

Causal Inference

This is the second post in a series of three on causality. In the last post I introduced this "new science of cause and effect" [1], and gave a flavor for causal inference and causal discovery. In

 <https://towardsdatascience.com/causal-inference-962ae97cefda>



- [1] [The Book of Why: The New Science of Cause and Effect](#) by Judea Pearl
- [2] Pearl, J. (2012). The Do-Calculus Revisited. [arXiv:1210.4852](#) [cs.AI]
- [3] Amit Sharma, Emre Kiciman. DoWhy: An End-to-End Library for Causal Inference. 2020. <https://arxiv.org/abs/2011.04216>
- [4] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [<http://archive.ics.uci.edu/ml>]. Irvine, CA: University of California, School of Information and Computer Science. <https://archive.ics.uci.edu/ml/datasets/census+income>
- [5] Künzel, Sören R., et al. "Metalearners for Estimating Heterogeneous Treatment Effects Using Machine Learning." *Proceedings of the National Academy of Sciences*, vol. 116, no. 10, Mar. 2019, pp. 4156–65. www.pnas.org, <https://doi.org/10.1073/pnas.1804597116>.