SFERS DATA Seminar with Python: 4

Yitae Kwon

SFERS of SNU

2024-1

Contents

Basic Concepts of Causal Inference

Recall: Regression Setup

$$y = X\beta + \epsilon$$

이고, 이들은 각각

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

입니다. 이때 우리는 y와 ϵ은 n차원 확률벡터로 바라보며, X와 β는 이미 **외생적으로 관측된** $n \times (k+1)$ 행렬과 정해진 계수의 $k+1$차원 벡터로 취급합니다. 차원을 계산해보면, 좌변과 우변이 모두 길이 n인 벡터가 됨을 알 수 있습니다.

Recall: Regression Setup

Fixed Design: **model matrix** 혹은 **design matrix** X가 non-random. 실험과 같이 우리가 X를 조정할 수 있다.
Random Design: (X_i, y_i)가 IID 분포를 따르게 random하게 관측된다. 단, 이 경우에도 X_i와 ϵ_i는 독립적이라고 가정한다.

Random Design 셋업 하에서도 $\hat{y}$, $\hat{\beta}$, $\hat{\sigma}^2$에 대한 추론은 동일하게 진행합니다. 단, 이 경우 이들은 모두 관측값으로부터 얻은 X에 **조건부**입니다. 이 때문에 X의 변동성으로 인한 추정량의 분산 증가 등이 발생할 수 있습니다.

이제 논의를 확장하여 보겠습니다. Fixed Design에서 Random Design으로의 확장은, X_i이 우리가 통제할 수 없는 내생변수인 상황을 생각하게 합니다. 관측연구처럼 말이죠. 이제 문제는 관측연구 상에서는 X_i와 ϵ_i의 독립성이 성립하지 않는 경우가 많다는 것입니다.

Causal Model

계량경제학 전반에서는 모형, 특히 회귀모형에 근거한 인과추론 방법론들이 주로 개발되어 왔습니다. 우리는 오늘 아래의 모형을 주로 사용할 것입니다.

$$Y = Z\beta + \epsilon \tag{1}$$

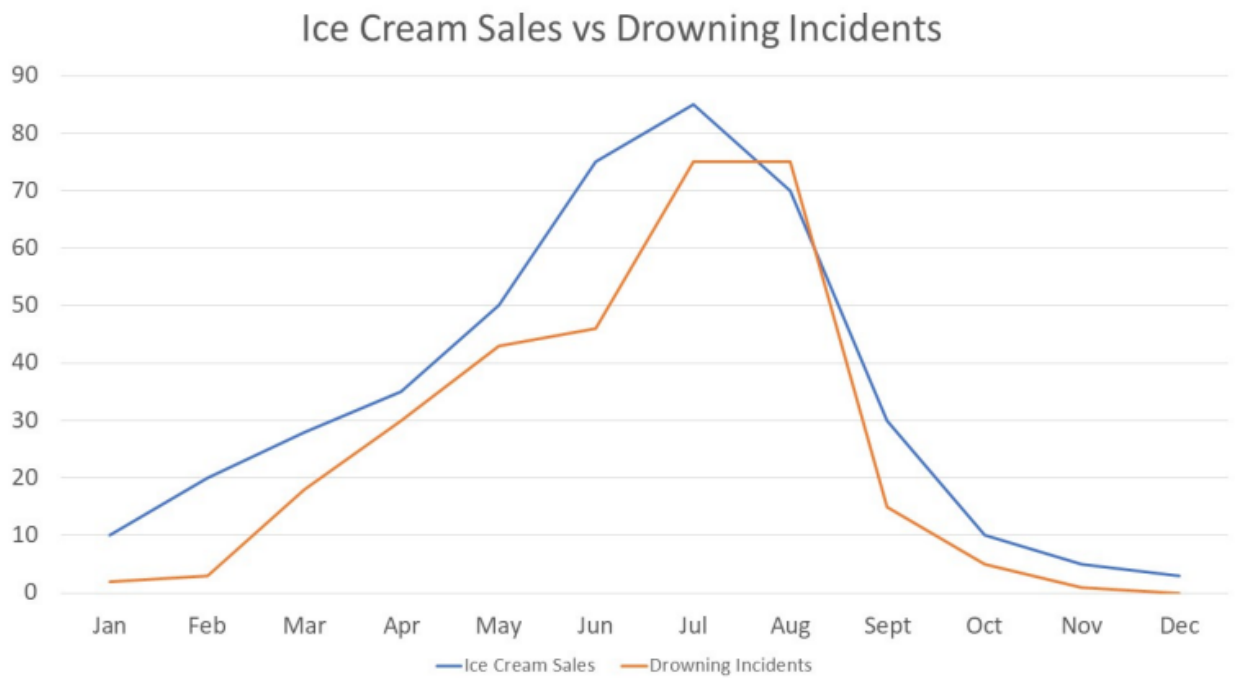
각 변수들은 아래를 의미합니다.

- Y는 outcome, 혹은 반응변수입니다. e.g. 실업률, 소득
- Z는 개체의 처리 여부/정도를 의미합니다. e.g. 경제성장률, 학벌
- β는 처리 효과를 의미합니다. e.g. Okun's coefficient
- ϵ은 오차를 의미합니다.

이제 경제성장률이 실업률에 미치는 영향을 파악할 수 있을까요?

Causality vs. Associativity

정답은 실험연구 상황에서는 O, 관찰연구 상황에서는 X입니다.
Example 1.



아이스크림의 판매량이 증가할수록 익사사고가 증가하는 경향성이 나타난다. 너무 맛있는 나머지 정신을 잃어버리는 게 아닐까?

Causality vs. Associativity

Example 2.

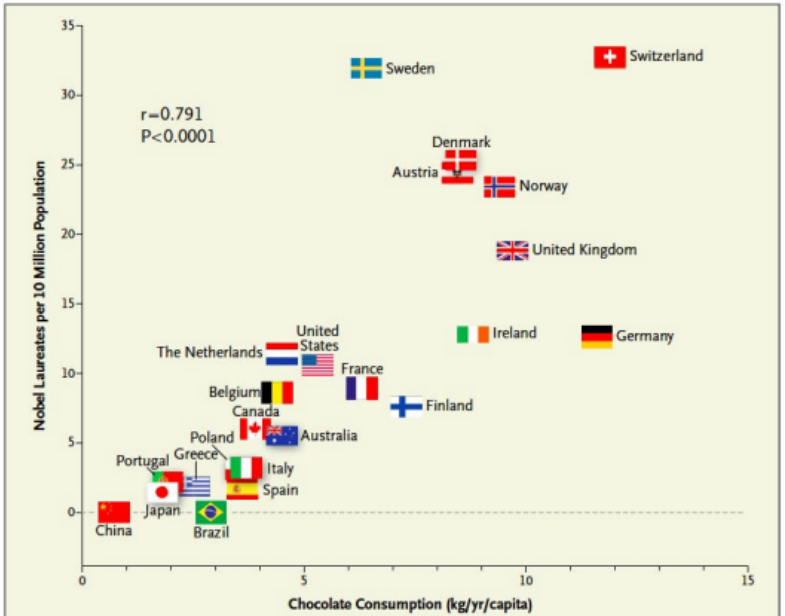


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

초콜릿 소비량이 많은 국가일수록 인구당 노벨상 수가 증가한다고 한다. 초콜릿이 연구 능력에 도움을 주는 게 아닐까? (Messerli, 2012)

Causality vs. Associativity

앞의 예시들에서 볼 수 있듯이 회귀분석이 만능은 아닙니다. 회귀분석은 두 변수 간의 **상관성(associativity)**를 확인하는 도구일 뿐, 어떤 변수가 다른 변수에 미치는 효과가 있는지를 다루는 **인과성(causality)**를 확인하는 도구는 아닙니다. 그 가장 큰 원인 중 하나는 Z와 ϵ에 존재할 수 있는 상관관계입니다. 우리는 이를 **endogeneity(내생성)**이라고 부릅니다. 내생성이 존재하는 모형이 가질 조건을 식으로 표현하면,

$$Z_i \not\perp\!\!\!\perp \epsilon_i$$

와 같습니다. 회귀분석 상황에서는 특히

$$\mathbb{E}[Z^T \epsilon] \neq 0$$

인 상황을 경계합니다.
Quiz: Z가 p개의 변수로 구성되면, $Z^T \epsilon$의 차원은?

Endogeneity in Regression

이러한 내생성은 회귀분석에서 왜 문제가 될까요? 먼저 다시 회귀모형 (1)으로 돌아가 보겠습니다.

$$Y = Z\beta + \epsilon$$

우리는 fixed design이든, random design이든, $\hat{\beta}$를 어떻게 구하냐면...

$$\hat{\beta} = (Z^T Z)^{-1} Z^T Y = (Z^T Z)^{-1} Z^T Z\beta + (Z^T Z)^{-1} Z^T \epsilon$$

로 구합니다. 따라서 우리는

$$\mathbb{E}[\hat{\beta}] = \beta + \mathbb{E}[(Z^T Z)^{-1} Z^T \epsilon]$$

을 얻습니다.

Endogeneity in Regression

fixed design에서는 Z가 fixed였으므로,

$$\mathbb{E}[(Z^T Z)^{-1} Z^T \epsilon] = (Z^T Z)^{-1} Z^T \mathbb{E}[\epsilon] = 0$$

을 쉽게 얻습니다. 즉 $\hat{\beta}$는 unbiased estimator입니다.
random design에서는 살짝 복잡하긴 하지만, Z와 ϵ이 독립임을 가정하므로,

$$\mathbb{E}[(Z^T Z)^{-1} Z^T \epsilon] = \mathbb{E}[(Z^T Z)^{-1} Z^T]\mathbb{E}[\epsilon] = \mathbb{E}[(Z^T Z)^{-1} Z^T] \cdot 0 = 0$$

입니다. 따라서 $\hat{\beta}$는 여전히 unbiased estimator입니다. 물론 계산하지는 않겠지만 fixed design에 비해 그 분산은 증가합니다.

Endogeneity in Regression

endogeneity가 발생하는 상황에서는 어떨까요?

$$\mathbb{E}[(Z^T Z)^{-1} Z^T \epsilon] \neq 0$$

입니다. 즉 $\mathbb{E}[\hat{\beta}] \neq \beta$입니다. 이는 $\hat{\beta}$가 편향추정량임을 의미합니다.

Note: 사실 정확히는 bias가 있을 조건이

$$\mathbb{E}[(Z^T Z)^{-1} Z^T \epsilon] \neq 0$$

로 주어집니다. 이는 무엇을 의미할까요?

$$\epsilon = Z\gamma + \eta$$

의 회귀계수 γ가 0인 것과 동일합니다.

Endogeneity in Regression

이는 당연한 것 아니냐구요? 회귀분석을 하면 ϵ에서 Z의 영향이 모두 제거되니 $\gamma \neq 0$일 수는 없지 않냐구요? 주의할 것은 위 모형은 실제 모형입니다. 우리가 찾은 $\hat{\beta}$는

$$\epsilon = Z0 + \eta$$

이게 만드는 회귀계수가 아니라, 이를 적합하여

$$\hat{\epsilon} = Z\hat{\gamma} + \hat{\eta}$$

처럼 표현했을 때 $\hat{\gamma} = 0$으로 만드는 회귀계수입니다. 즉 $\hat{\epsilon}_i$는 Y를 Z들의 선형결합으로 만들어지는 linear space에 projection시킨 뒤 남은 residual vector들이기에 Z_i와 sample에선 (linearly) independent하지만, 그것이 실제 poplulation에서 ϵ과 Z의 독립성을 의미해주진 못합니다.

Endogeneity in Regression

지난 시간에 다루었던 통계적 추정이론을 다시 봅시다.

- **bias(편향)**: $\text{bias}(\hat{\beta}) = \mathbb{E}[\hat{\beta}] - \beta$
- **MSE(평균제곱오차)**: $\text{MSE}(\hat{\beta}) = \mathbb{E}[(\hat{\beta} - \beta)^2] = \text{Var}(\hat{\beta}) + (\text{bias}(\hat{\beta}))^2$

일반적으로 우리는 MSE처럼 제곱합을 최소화시키는 추정량들을 선호합니다(회귀분석의 추정량이 얻어지는 과정과 마음을 같이 합니다.). MSE는 추정량의 분산과 편향의 제곱의 합으로 분해할 수 있습니다.

이때 일반적으로 표본의 개수 n이 증가할수록 $\text{Var}(\hat{\beta})$는 감소하는 반면, $\text{bias}(\hat{\beta})$는 그렇지 않습니다. 따라서 일반적으로 통계적 추정방법을 평가할 때에는 편향이 0인 불편추정량을 선호합니다. 불편추정량들끼리는 그 분산을 비교하여 추정의 **효율성**(efficiency)를 따지게 됩니다. 추정량의 분산이 감소할수록 신뢰구간의 길이가 줄어들고 실제 참값과 더 가까운 추정량을 얻을 확률이 증가하기에, 더 효율적인 추정량이라고 부릅니다.

Endogeneity in Regression

MSE가 0으로 수렴한다는 것은 무슨 의미일까요? 바로 n이 증가함에 따라 $\hat{\beta}$가 β로 수렴하게 된다는 것입니다.

만약 모든 $\epsilon > 0$에 대하여,

$$\lim_{n \to \infty} P(|\hat{\beta} - \beta| > \epsilon) = 0$$

이라면, $\hat{\beta}$는 β로 **확률수렴**한다고 말하며, 추정량 $\hat{\beta}$는 **일치추정량**(consistent estimator)가 됩니다. 일치추정량이 되기 위해서는 편향과 분산이 모두 0으로 수렴해야만 합니다.

그런데 endogeneity 하에서 $\hat{\beta}$는 편향이 존재하므로, 일치추정량이 아닙니다. 즉 표본 크기가 아무리 증가해도 우리는 제대로 된 효과를 얻지 못하게 됩니다.

Source of Endogeneity

Example 1. Omitted Variable Bias
아까 저희가 endogeneity에 관해 이야기하면서

$$\epsilon = Z\gamma + \eta$$

의 실제 회귀계수 γ가 0이 아닐 때 문제임을 밝혔습니다. 다르게 말하면 오차 ϵ 부분에 Z와 실제로 관련이 있는 변수가 포함되어 있을 때 문제가 발생합니다. 따라서 실제 모형이

$$Y = Z\beta + W\delta + \xi$$

이고 $(Z, W) \perp\!\!\!\perp \xi$라면 원래 모형에는 내생성이 없습니다. 그러나 우리가 W를 누락해 버린다면,

$$Y = Z\beta + (W\delta + \xi) = Z\beta + \epsilon$$

을 적합하게 됩니다.

Example 1. Omitted Variable Bias
그렇다면

$$\begin{aligned}\mathbb{E}[(Z^T Z)^{-1} Z^T \epsilon] &= \mathbb{E}[(Z^T Z)^{-1} Z^T (W\delta + \xi)] \\ &= \mathbb{E}[(Z^T Z)^{-1} Z^T W]\delta\end{aligned}$$

이고, 우리는 편향의 두 가지 인자를 알게 됩니다.

- bias의 크기는 누락된 변수 W의 계수 δ와
- W가 가진 정보 중 Z에 의해 설명될 수 있는 부분

에 편향이 비례합니다.

Example 2. Measurment Error
측정 결과에 오차가 존재할 경우에도 마찬가집니다. 이번에는 모형

$$Y = Z\beta + \xi$$

은 맞지만, Z가 연속형 변수라서 측정하는 과정에서 Z 대신 measurment error W가 포함된 $\tilde{Z} = Z + W$를 관측하게 된다고 합니다. 예를 들어 장학금 관련 설문조사를 할 경우 응답자들은 실제 소득 Z보다 의도적으로 낮게 기입하여 $\tilde{Z}$를 얻게 될 것입니다. 이 경우 W는 음의 경향성을 가집니다. 그렇다면 우리가 적합하는 모형은

$$Y = Z\beta + \xi = (\tilde{Z} - W)\beta + \xi = \tilde{Z}\beta + (-W\beta + \xi) = \tilde{Z}\beta + \epsilon$$

입니다. 유사한 방식으로 편향의 존재를 보일 수 있습니다.

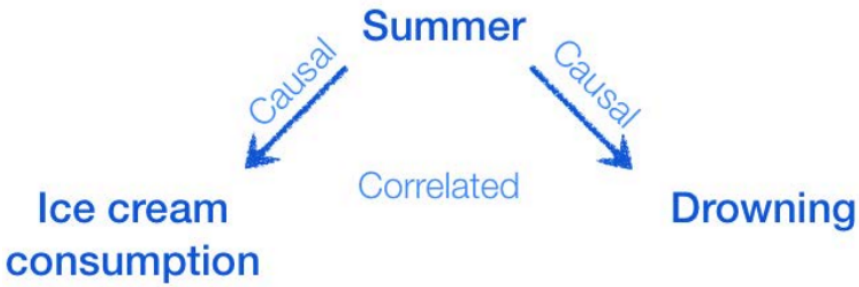
Example 3. Simultaneous Equations Bias
IS-LM 모형을 단순화시켜 아래와 같이 생각합니다. 편의상 통화량과 가격수준은 고정입니다.

$$Y = \alpha_0 - \alpha_1 i + u_{IS} \quad \text{(IS 곡선)}$$
$$M/P = \beta_1 Y - \beta_2 i + u_{MD} \quad \text{(LM 곡선)}$$

그렇다면 우리가 실물시장에서 이자율에 대한 산출량의 탄력성을 확인하고자 한다면, α_1가 추정하고자 하는 대상이 됩니다. 그렇다면 그냥 Y에 대해 i를 회귀분석하면 될까요? 답은 당연히 아니오입니다. 실제 데이터에서 얻은 Y와 i는 두 곡선의 교차점에서 만들어지며, 각 상황마다 α_0, M/P와 같은 변수들의 값이 변하게 (i, Y)의 데이터는 항상 어떠한 직선 위에 있는 것이 아닙니다.

Source of Endogeneity

Example 4. Confounding Effects



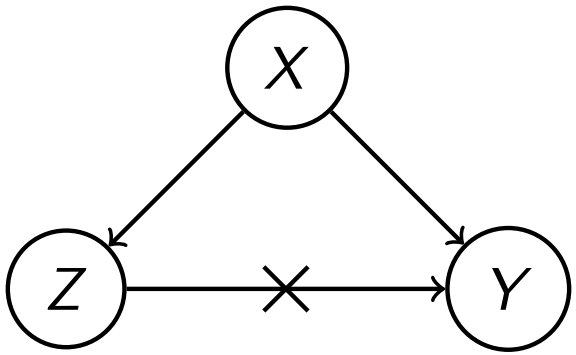
여름, 혹은 뜨거운 기온은 모두 인과적으로 아이스크림 소비량과 익사사고 수를 증가시킵니다. 이처럼 보려고 하는 두 요인에 모두 인과적으로 영향을 미쳐 두 요인 사이의 상관관계를 만드는 요인을 **교락요인**(confounder)라고 부릅니다. 마찬가지로 초콜릿 소비량과 노벨상 수상자 비율 간에는 '유럽'이라는 지역적 원인과 '소득' 등이 교락요인이 될 수 있습니다.

Example 4. Confounding Effects

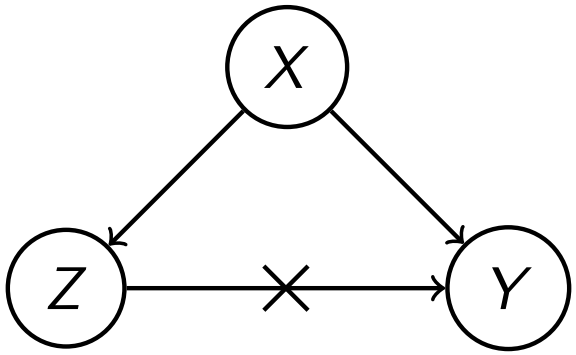
confounding effect는 endogeneity를 발생시키는 가장 주요한 요인 중 하나입니다. 수식으로 쓰면, 우리의 회귀모형

$$Y = Z\beta + \epsilon$$

에서 Z와 Y에 모두 (직접적인) 인과적 영향을 미치는 X를 고려할 수 있습니다. 주의할 것은 Z와 Y는 어떠한 인과관계도 가지지 않습니다.



Example 4. Confounding Effects

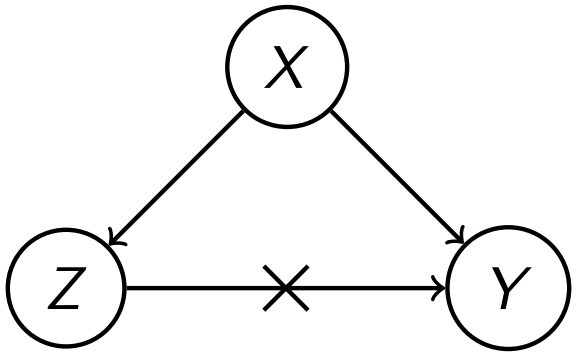


이제 Y를 반응변수로, Z를 설명변수로 하는 회귀모형을 고려하면

$$Y = Z\beta + \epsilon$$

입니다. 주의할 것은 여전히 이 회귀모형은 Z와 Y의 선형 상관성을 고려하는 데는 유의합니다. 어찌하였든 X에 의하여 Z와 Y가 유사한 경향성을 가진다고 보고하게 될 것이고, 이는 상관관계와 예측 측면에서는 아무런 문제가 없습니다.

Example 4. Confounding Effects

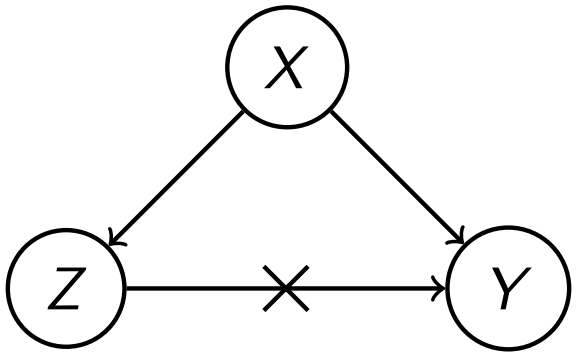


회귀분석을 수행하면, 인과적 효과는 0이기에 우리는 $\mathbb{E}[\hat{\beta}] = 0$이길 기대합니다. 그러나... 정말 단순하게 $Z = X$, $Y = X\xi$로 결정될 때를 생각하면

$$\mathbb{E}[\hat{\beta}] = \mathbb{E}[(Z^T Z)^{-1} Z^T Y] = \mathbb{E}[(X^T X)^{-1} X^T X\xi] = \xi \neq 0$$

일 수 있습니다.

Example 4. Confounding Effects

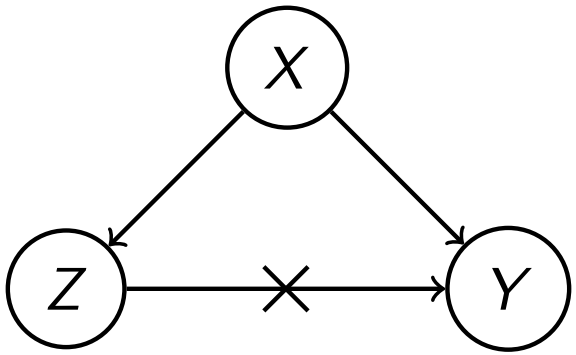


인과추론 상황에서 가장 많이 접하는 문제 중 하나는 어떠한 처리 Z가 Y에 미치는 영향을 볼 때의 교락 요인 문제입니다. 국비 지원 부트캠프(Z)의 청년 고용 촉진 효과(Y)를 확인하고자 합니다. 그렇다면 $Z \in \{0,1\}$, $Y \in \{0,1\}$인 이항변수입니다. 사실 지금도 많은 사람들이

$$Y = Z\beta + \epsilon$$

을 적합하여 얻은 β가 양수 p일 경우, 부트캠프가 청년 고용률을 $100p\%$만큼 높인다고 해석하고 있습니다.

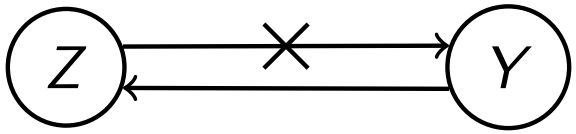
Example 4. Confounding Effects



여기에서 가능한 교락요인은 무엇일까요? 바로 구직자의 구직의지(X)가 될 수 있습니다. 구직의지가 높은 사람은 부트캠프에 더 많이 참여하려 합니다. 동시에 구직의지가 높은 학생은 부트캠프에 참여하든 말든 언젠가 취업할 확률이 높습니다. 이에 따라 실제로는 부트캠프의 고용촉진 효과가 없을지라도, 양의 상관관계는 존재하며, 회귀분석의 결과는 양의 상관성을 나타내게 됩니다.

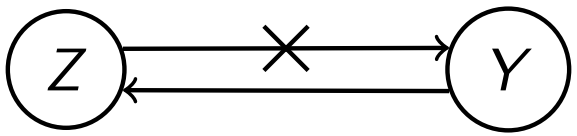
Source of Endogeneity

Example 5. Reverse Causality
만약 부트캠프-취업률 예시처럼 선후관계가 명확하다면, Z가 Y를 cause 하는지만 생각해보면 됩니다. 그러나 우리가 특정 시점의 자료만 가지고 있다면, 어떤 변수가 선행되고 다른 변수에 영향을 미치고 있는지 확인하기 어렵습니다.



이처럼 우리가 보고자 하는 관계에는 인과성이 없고, 오히려 반응변수가 설명변수에 영향을 미치는 상황을 **역인과관계**가 있다고 말합니다.

Example 5. Reverse Causality



여러 국가들에서 노동인권(Z)와 노조 조직 정도(Y)를 관측한 뒤 회귀분석을 돌려봤다고 합시다. 실제로는 노조가 많이 조직된 국가, 즉 Y가 큰 국가일수록 그들의 노력으로 노동인권 Z가 올라간 상황입니다. 그러나 친기업 연구자들은 오히려 '이넘들이 노동인권이 높으니까 배가 불러서 귀족노조를 더 조직하려 한다'라고 결론을 내릴 수도 있겠습니다.

Source of Endogeneity

Example 6. Regression to the Mean
서울시가 학생인권조례를 폐지한 것을 두고 진보교육계에서는 반대 의견이 뜨겁습니다. 반면 보수성향 교육감들은 진보교육감이 제정한 학생인권조례가 비가역적인 교육시스템의 손상을 가져왔다며, 지금이라도 폐지한 것을 다행으로 여기는 분위기입니다. 사실 2010년 학생인권조례가 제정되었을 때만 해도, 보수교육감과 보수성향 교원들은 체벌을 금지하면 면학 분위기 조성이 어렵다며 이를 반대했었습니다. 체벌이 과연 성적 향상에 도움이 될까요?

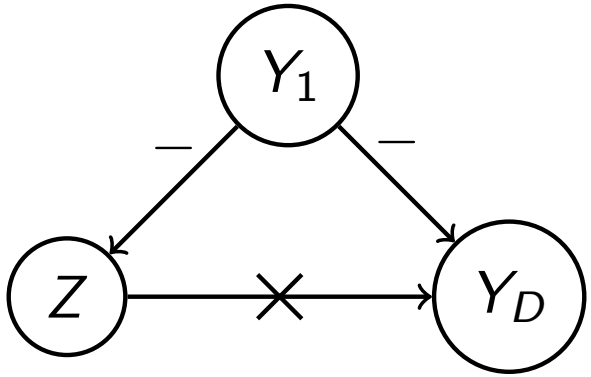
이제 우리의 상황을 생각해 보겠습니다. 중간고사 때 학생들의 성적 Y_1을 보고, 하위 그룹에는 체벌을 가하고, 상위 그룹에는 체벌을 하지 않았습니다. 그 다음 체벌 여부(Z)와 기말고사 성적 상승량($Y_2 - Y_1$)를 취합하여 연줄이 있는 보수교육감에게 넘겼습니다. 그 다음 $Y_2 - Y_1$를 반응변수로, Z를 설명변수로 하여 회귀분석을 했더니 양의 회귀계수를 얻어 체벌이 성적 상승에 효과가 있음을 주장해버렸습니다. 문제가 무엇일까요?

Example 6. Regression to the Mean
정답은 역시 내생성에 있습니다. 편의상 모든 학생들이 모두 동일한 학생이며, 체벌의 효과는 없다고 하겠습니다. 그렇다면 평균 μ에 대하여

$$Y_1 = \mu + u_1$$
$$Y_2 = \mu + u_2$$

처럼 Y_1, Y_2는 평균으로부터 단지 적당한 변동만 있는 값입니다. μ_1과 μ_2는 독립적으로 결정됩니다. Y_1이 낮은 학생들에게 처벌을 가한다면, 낮은 Y_1, 혹은 낮은 u_1은 Z를 유도합니다. 동시에, $Y_D = Y_2 - Y_1 = u_2 - u_1$은 상승하는 효과를 가지게 됩니다.

Example 6. Regression to the Mean
이를 그림으로 그리면 아래와 같습니다.



따라서 이 경우에는 우리가 분석하고자 하는 시점 이전에 결정된 Y_1이 교락 요인으로 작용하여 Z와 Y_D에 모두 음의 인과관계를 가지게 되고, 이에 따라 Z와 Y_D에는 실제로 아무런 인과성이 없음에도 양의 상관관계가 등장하는 것입니다.

Example 7. Simpson's Paradox
SFERS의 인기가 너무 높아져 SFERS를 대비하는 스터디가 생긴 미래를 고려해 봅시다. 어떤 학생은 과연 이 스터디가 SFERS 합격에 도움이 되는지 분석해보고자, 2개 학부의 학생들에 대하여 SFERS 합격 여부와 SFERS 대비 스터디 참여 여부를 모아 봤습니다.

합격자수/지원자수	스터디 참여	스터디 불참	총계
경제학부	20/50	50/200	40/250
김치볶음밥학부	10/100	1/20	11/120
총계	30/150	51/220	

그리고 실수로 설문조사 과정에서 학부를 누락해 버려서, 스터디 참여자들($Z = 1$)인 사람 150명 중 20퍼센트인 30명이 합격했으며, 스터디 불참자($Z = 0$) 220명 중 23퍼센트인 51명이 붙었다고 확인했습니다. 그래서 SFERS 대비 스터디에 불참하고 말았습니다...

Source of Endogeneity

Example 7. Simpson's Paradox

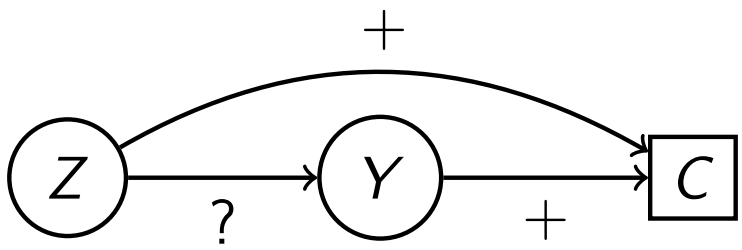
합격자수/지원자수	스터디 참여	스터디 불참	**참여 비율**
경제학부	20/50	50/200	50/250
김치볶음밥학부	10/100	1/20	100/120
총계	30/150	51/220	

그러나 학부를 나누어서 보면, 경제학부의 경우 스터디 불참시 합격률이 25퍼센트 선에서 40퍼센트까지 올라가며, 김치볶음밥학부 역시 5퍼센트 선에서 10퍼센트 선까지 상승합니다. 즉 스터디는 사실 합격에 도움이 됩니다. 이 문제는 omitted variable bias로 설명됩니다. 학부는 생략된 변수로써 작용하는데, 스터디 참여 여부 Z와 경제학부인지 여부는 음의 상관관계를 맺습니다. 반면 경제학부라면 SFERS 합격과 양의 상관관계가 있을 것입니다. 이에 따라 참여 여부와 합격 여부에는 실제론 양의 직접적 인과관계가 있음에도, 회귀분석의 결과는 학부에 의한 음의 상관관계가 이를 지배하여 음의 값이 나오게 되는 것입니다. 이를 제거하기 위해서는 학부를 회귀모형에 넣어 그 영향을 분리해 주어야 합니다.

Source of Endogeneity

Example 8. Selection Bias

취업대비 직무교육/영어교육 프로그램 Z에 참여한 사람들의 영어 점수 Y를 확인하고자 합니다. 단, 취업에 성공한 사람들($C = 1$)만 여기에 응답하였으며, 아직 구직하는 사람들은($C = 0$)은 미응답하였습니다.



Z와 Y의 상관관계는 두 경로에 의해 만들어집니다.

- $Z \to Y$의 직접적인 인과관계.
- $Z \to C \leftarrow Y$의 C로 인한 conditioning에서 발생하는 상관관계.

이처럼 응답자가 어떠한 특성을 가지고 있는 사람들에만 한정되는 경우 $C = 1$인 사람들에 조건부로 Z와 Y의 관계를 파악하게 되므로, 둘째 경로에 의한 bias가 생겨 우리가 원하는 첫째 경로를 정확히 알 수 없습니다. 이를 **selection bias**(선택 편향)이라 합니다.

Potential Outcome Framework

이러한 문제들이 곳곳에 도사리고 있기 때문에, 회귀분석 결과를 상관관계가 아닌 인과관계로 섣불리 해석하는 것은 금지됩니다. 그러나 그 해석이 가능한 상황이 있습니다. 바로 모든 것의 근원인 endogeneity를 제거하는 것입니다. 즉

$$Z \perp\!\!\!\perp \epsilon$$

을 보장할 수 있는 상황을 고려하는 것입니다. 이처럼 독립성이 만족되도록 하여 인과적 효과를 **식별**(identify)할 수 있게 만드는 과정을 연구의 **디자인(design)**이라고 부릅니다. 이러한 디자인은 비용, 물리적 가능성, 해당 분야의 관습 등을 고려하여 만들어져야 합니다. 따라서 통계학자가 제작한 디자인에 기대는 것이 아니라, 해당 분야의 전문가(경제학자, 의학자 등)이 주도적으로 디자인해야만 리서치 퀘스천에 맞는 적절한 분석을 수행할 수 있습니다.

Potential Outcome Framework

Potential Outcome Framework는 리서치 퀘스천과 문제 상황에 맞는 적절한 인과적 디자인을 묘사할 수 있는 도구입니다. 인과추론에서 가장 문제가 되는 것은, 처리군의 개체가 처리를 안 받았을 때의 결과값, 그리고 대조군의 개체가 처리를 받았을 때의 결과값을 절대 관측할 수 없다는 것입니다. 우리는 실제 관측된 결과의 반대 상황을 상상해볼 수 있습니다. 이를 수학적으로 묘사하려는 시도가 potential outcome framework입니다. 먼저, 우리의 연구에 n명의 피험자가 들어와 있다고 합시다.

- Outcome: $Y_1, Y_2, \cdots, Y_n$
- Treatment: $Z_1, Z_2, \cdots, Z_n$

을 우리는 관측합니다. 여기에서는 잠시 Z_i를 처리를 받지 않았으면 0, 처리를 받았으면 1인 이항변수로 생각하겠습니다.

Potential Outcome Framework

이제 i번째 피험자가 처리를 받았을 때($Z_i = 1$)의 관측값과 처리를 받지 않았을 때의 potential outcome을 각각

$$Y_i(1), Y_i(0)$$

으로 쓸 수 있으며, 실제로 관측된 값 Y_i는

$$Y_i = Y_i(Z_i) = Z_i Y_i(1) + (1 - Z_i) Y_i(0)$$

으로 주어지게 됩니다. 다르게 말하면, $Y_i(1 - Z_i)$는 절대로 관측할 수 없습니다.

Potential Outcome Framework

그렇다면 individual level에서 i번째 피험자에게 주어진 처리의 인과적 효과는

$$\beta_i = Y_i(1) - Y_i(0)$$

으로 나타납니다. 주의할 것은 $Y_i(1), Y_i(0)$은 개체가 고정된다면 고정된 값이라고 취급한다는 점입니다. 따라서 β_i도 i가 고정되면 고정입니다. 단 전체 모집단의 관점에서는 각 개인 i가 서로 다른 성질을 가지므로, $Y_i(1), Y_i(0), \beta_i$를 $Y(1), Y(0), \beta$에 대한 i번째 표본이라고 생각해줄 수도 있습니다. 단 β_i는 Z_i에 의해서는 절대 변하지 않습니다. 또한 β_i는 절대 추정할 수 없는 값입니다. 우리는 평행세계를 보지 않는 한 $Y_i(1 - Z_i)$ 부분을 절대 관측할 수 없고, 이에 따라 β_i는 절대 얻을 수 없는 값이 됩니다.

Potential Outcome Framework

이에 따라 많은 인과추론 연구에서는 β_i들의 평균을 추정함으로써 처리 효과를 확인하고자 합니다. 통계학에서 추정하고자 하는 모수들을 **estimand**라 부르며, 아래 세 개는 주요한 **causal estimand**입니다.

- **ATE(Average Treatment Effect)**

$$ATE = \mathbb{E}[Y(1) - Y(0)]$$

- **ATT(Average Treatment effect on the Treated)**

$$ATT = \mathbb{E}[Y(1) - Y(0)|Z = 1]$$

- **ATC(Average Treatment effect on the Control)**

$$ATT = \mathbb{E}[Y(1) - Y(0)|Z = 0]$$

Potential Outcome Framework

이들은 서로 그 값이 다를 수 있습니다. 포인트는 어떠한 교락요인의 존재로 처리에 따른 효과의 크기와 처리 여부 사이에 상관성이 존재할 수 있다는 것이다.
어떠한 누락된 변수 X_i이 $U(0,1)$을 따르고,

$$Y_i(1) = 5X_i, \quad Y_i(0) = 3X_i, \quad Z_i = \begin{cases} 1 & X_i > 0.5 \\ 0 & X_i \leq 0.5 \end{cases}$$

처럼 결정된다고 해 봅시다. 그렇다면

$$\begin{aligned} ATE &= E[5X - 3X] = 1 \\ ATT &= E[5X - 3X | Z = 1] = 1.5 \\ ATC &= E[5X - 3X | Z = 0] = 0.5 \end{aligned}$$

로 상이함을 알 수 있습니다. 위의 예시를 조금 더 극단적으로 만들면, 처리 효과의 부호가 완전히 바뀌어 버릴 수도 있습니다.

Potential Outcome Framework

그렇다면 어떠한 estimand를 사용해야 할까요? 이는 리서치 퀘스천, 그리고 실험 디자인에 달려 있습니다. 예를 들어, 어떠한 정책을 수행하고 그 정책영향을 평가하고자 합니다. 이는 처리가 이루어진 후 처리받은 개체들이 얼마나 효과를 봤는지 판단하는 후향적(retrospective)인 연구입니다. 이는 처리를 받은 개체들의 처리 효과를 보므로, ATT를 추정할 수 있는 디자인을 사용해야 합니다.

한편 추가적인 정책 시행을 하였을 때 아직 처리를 받지 않은 피험자들이 어떠한 영향을 받을지 궁금할 수 있습니다. 예비타당성조사와 같은 상황이 여기에 해당합니다. 이는 전향적(prospective)한 연구이며, 처리를 받지 않은 개체들의 처리 효과를 봐야 하므로 ATC를 추정할 수 있는 디자인을 사용해야 합니다.

더 나아가서는 단지 처리 효과가 전체 집단에 어떠한 영향을 평균적으로 미치는지 궁금할 수 있습니다. 자연과학의 상황이 여기에 많이 해당합니다. 이 경우 ATE를 추정해야 합니다.

Fundamental Problem of Causal Study

개체 i	처리여부 Z_i	소득 $Y_i(1)$	$Y_i(0)$
1	1	2000	1500
2	0	2500	2100
⋮	⋮	⋮	⋮
n	1	6000	3500

우리가 원하는 테이블은 위처럼 각 개체의 처리여부와 두 potential outcome을 모두 가지고 있는 테이블입니다. 이 정보가 있으면 우리는 쉽게 ATE, ATT, ATC를 모두 추정할 수 있습니다.

Fundamental Problem of Causal Study

개체 i	처리여부 Z_i	소득 $Y_i(1)$	$Y_i(0)$
1	1	2000	????
2	0	????	2100
⋮	⋮	⋮	⋮
n	1	6000	????

그러나 우리가 가진 테이블은 위와 같습니다. 즉 $Y_i(1 - Z_i)$ 부분은 관측할 수 없습니다. 이제 Z_i가 연속변수라거나 3개 이상의 처리가 존재한다면, 이러한 문제는 더욱 심각해집니다. 우리의 목표는 적절한 방식으로 $Y_i(1 - Z_i)$를 잘 채워넣어 $\hat{\beta}_i$, 혹은 그 평균을 추정해내는 것입니다.

$$\hat{\beta}_i = \begin{cases} Y_i(1) - \hat{Y}_i(0) & (Z_i = 1) \\ \hat{Y}_i(1) - Y_i(0) & (Z_i = 0) \end{cases}$$

Fundamental Problem of Causal Study

그러나 정확히 $\hat{\beta}_i$를 알 필요는 없습니다. 우리는 그 평균을 추정하는 것이기 때문에, $\hat{Y}_i(1), \hat{Y}_i(0)$들을 모두 알 필요 없이

$$\hat{\mathbb{E}}[Y_i(1)], \hat{\mathbb{E}}[Y_i(0)]$$

만 추정해도 됩니다. 앗! 그런데 우리는 어떠한 대상의 평균을 추정하는 좋은 방법을 알고 있습니다. 바로 회귀분석입니다. 이 때문에 우리는 인과적 효과를 보기 위해 Z_i에 의해 인과적으로 바뀌는 Y_i의 회귀모형을

$$Y_i = Y_i(Z_i) = \alpha + Z_i\beta + \epsilon_i$$

처럼 만들어 주어진 Z_i와 Y_i 정보를 적합시키고, β를 추정해왔습니다. 또한

$$\hat{\mathbb{E}}[Y(1)] = \hat{\alpha} + \hat{\beta}$$
$$\hat{\mathbb{E}}[Y(0)] = \hat{\alpha}$$

으로 결정되므로, $\hat{\beta}$가 ATE의 추정량이 됩니다.

Revisit Endogeneity

우리는 이제 회귀분석이 왜 인과추론에서 자주 사용되는지를 potential outcome framework을 통해 확인하였습니다. 그러나 여전히 문제는 회귀모형

$$Y_i = \alpha + Z_i\beta + \epsilon_i \tag{2}$$

이 β를 제대로 추정할 수 있냐는 것입니다. 정확히는 *ATE*의 추정량 $\hat{\beta}$의 consistency와 unbiasedness가 보장될 수 있냐는 것입니다.

여전히 내생성 문제는 남아 있습니다. 왜냐하면, 우리가 $Y_i(Z_i)$를 모형화하기 위해 사용한 식에서는 Z_i와 ϵ_i의 상관관계가 존재할 수 있기 때문입니다. 우리가 potential outcome framework를 이용할 때에 각 i에 대해서는 ϵ_i가 고정되어 있음에 따라 Z_i가 바뀜에 따른 효과가 β로 나타나지만, 회귀분석 과정에서는 ϵ_i에 존재하는 Z_i의 상관성이 모두 $\hat{\beta}$로 흡수되어 나타나기 때문입니다.

Revisit Endogeneity

이제 potential outcome framework 하에서 이를 해결하는 방법을 생각하여 보겠습니다. potential outcome framework에서 Z_i와 ϵ_i의 상관성은 Z_i가 주어지는 **assignment mechanism**으로부터 등장한다고 생각합니다. 즉 ϵ_i가 높은 개체 i가 높은 처리 확률을 가지게 된다면 둘 사이에는 양의 상관관계가 존재하게 됩니다. 이를 끊는 방법 중 하나는 Z_i가 개체의 특성에 완전히 무관하게 만드는 것입니다. 이처럼 Z_i를 실험자가 무작위로 조정함으로써 둘 사이의 상관관계를 없애는 디자인을 **랜덤화 실험**(randomized controlled trials; RCT)이라고 부릅니다.

랜덤화 실험에서는 Z_i를 완전히 무작위로 결정합니다. 즉 각 개체에 대하여 처리 여부를 동전 던지기를 통해 결정하는 상황을 고려해볼 수 있습니다. 그렇다면 Z_i는 당연하게도 ϵ_i와 무관하게 되며, 내생성이 없고, $\hat{\beta}$는 ATE의 불편추정량이 됩니다.

Revisit Endogeneity

애초에 RCT에서는 회귀분석을 수행할 필요도 없습니다. 이 경우 회귀분석의 결과에서 얻는 값은

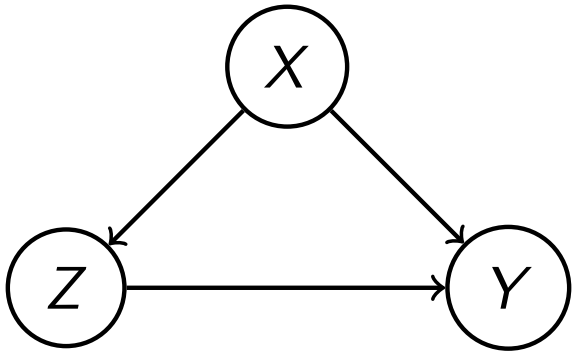
$$\hat{\beta} = \frac{1}{n_t}\sum_{i=1}^{n} Z_i Y_i - \frac{1}{n_c}\sum_{i=1}^{n}(1 - Z_i) Y_i$$

와 동일하며, 단지 표본평균의 차일 뿐입니다. 이는 내생성의 부재로부터 등장하게 됩니다.

그러나 **관찰연구**(observational study)에서는 우리가 Z_i를 조정하기 어려우므로, Z_i과 ϵ_i의 상관관계를 0으로 만드는 Z_i를 결정할 수 없게 됩니다. 이에 따라 내생성을 없애는 다른 접근법이 필요하게 됩니다.

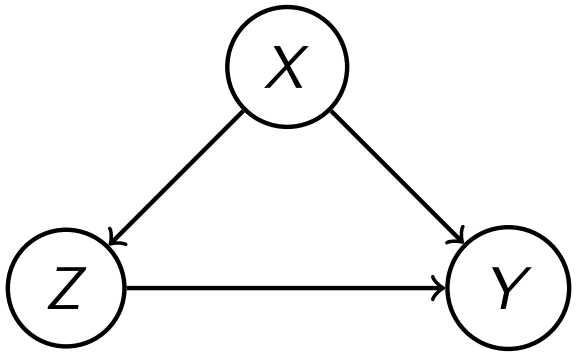
Revisit Endogeneity

관찰연구에서 사용될 수 있는 방법은 Z_i와 ϵ_i 사이의 상관성을 파악하고 이를 보정해 주는 것입니다. Z_i의 assignment mechanism이 개체 i의 개별적인 성질인 **공변량** (covariates) X_i에 의해 결정된다고 생각해 보겠습니다. 그리고, ϵ_i, 즉 $Y_i(1)$과 $Y_i(0)$의 전반적인 수준을 결정하는 값 역시 X_i에 의해 결정된다고 생각하겠습니다. 그렇다면 이를 아래처럼 그릴 수 있습니다.



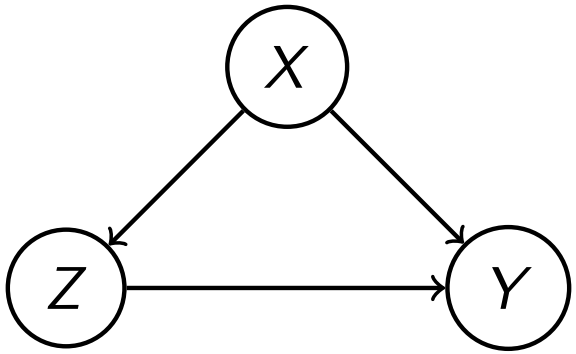
이제 우리의 목표는 $Z \to Y$ 경로의 영향을 확인하기 위하여, $X \to Z$와 $X \to Y$로부터 $Z \leftarrow X \to Y$ 경로로 인한 상관성을 통제하는 것입니다.

Revisit Endogeneity



이제 그 통제를 위하여 관찰연구에서의 많은 디자인이 개발되었습니다. 오늘은 그 방법 중 **도구변수법**(Instrumental Variable Method)를 집중적으로 알아볼 것입니다. 한편 특정한 디자인과 가정 하에서는 그 관계를 자동으로 제거해줄 수 있습니다. 이러한 디자인을 **준실험**(quasi-experiment) 혹은 **자연실험**(natural experiment)라고 부릅니다. 우리는 이후 이러한 디자인이 어떤 가정 하에서 만들어지며, 어떻게 저 상관관계를 통제하는지 확인해볼 것입니다.

Revisit Endogeneity



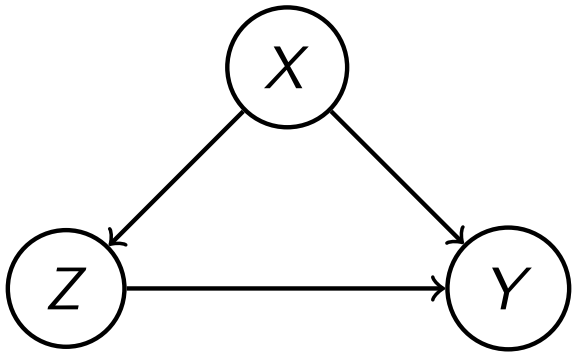
한편 이처럼 상관관계의 통제를 위하여 위처럼 변수 간의 인과관계를 causal graph의 형태로 그릴 수 있습니다. 이처럼 각 node(변수)들 사이의 edge(인과관계)가 단방향이면서, 순환 관계를 이루지 않는 그래프를 **DAG**(Directed Acyclic Graph)라 하며, 연구 디자인에서 자주 사용됩니다. DAG는 통계적으로 추정하기 쉽지 않기에, 전통적인 인과추론에서는 선험적인 지식을 바탕으로 이를 그린 후 그 위에서 추론을 수행합니다. 반면 최근에는 causal discovery 분야가 발달하고 있어, 이러한 구조적 종속성 자체를 추정하는 연구도 활발히 진행되고 있습니다.

Instrumental Variable Method

Instruments or Encouragement

우리가 앞서 경계한 내생성은 공변량 X가 Z와 Y에 confounder로 작용하여 나타났습니다. 이러한 관계를 제거하는 방법에는 무엇이 있을까요? 한 번 아래의 상황을 고려해 보겠습니다.

우리는 금연 여부(Z)에 따른 폐암 진행 정도(Y)를 확인하고자 합니다. 그러나 금연 여부를 우리가 마음대로 랜덤화하기는 어려운 노릇입니다. 그러나 생활 환경, 임금 수준, 교육 수준과 같은 공변량들은 confounder로 작용하여 회귀분석 결과를 왜곡할 수 있습니다. 흡연이 폐암 진행에 미치는 영향을 어떻게 계량화할 수 있을까요?



Instruments or Encouragement

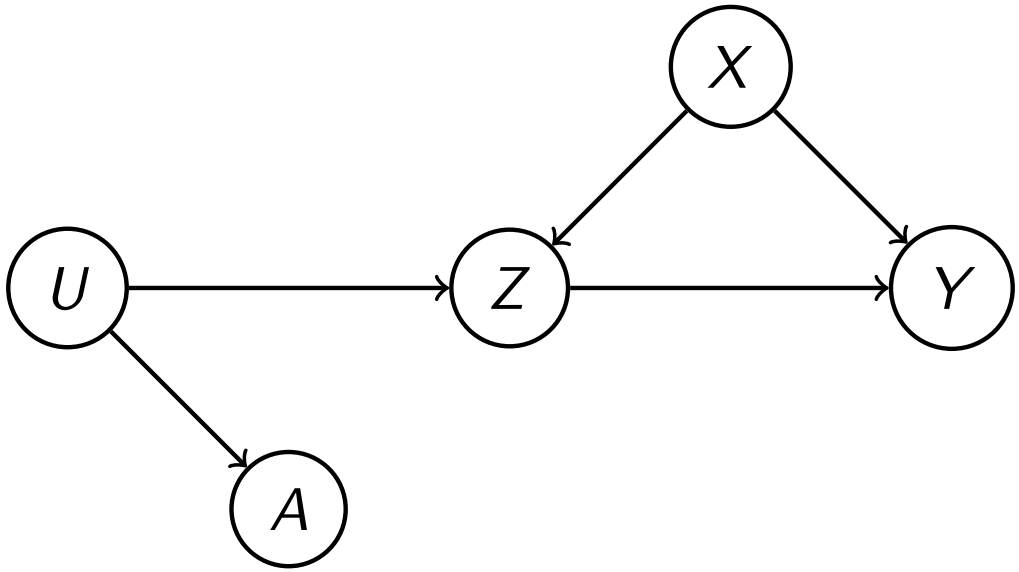
도구변수는 아래 세 조건을 만족하는 변수 A를 의미합니다.

(i) A는 Z와 상관관계를 가진다.

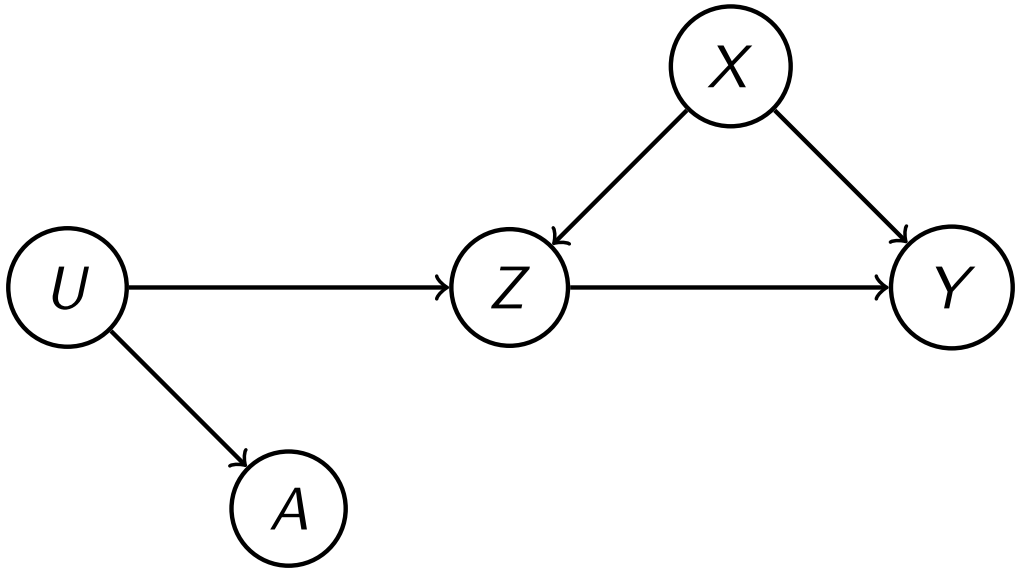
(ii) A는 Y에 직접적인 인과효과를 가지지 않는다. 즉 A의 Y에의 효과는 Z를 거치는 간접효과만 있다.

(iii) A와 Y는 동일한 원인을 공유하지 않는다.

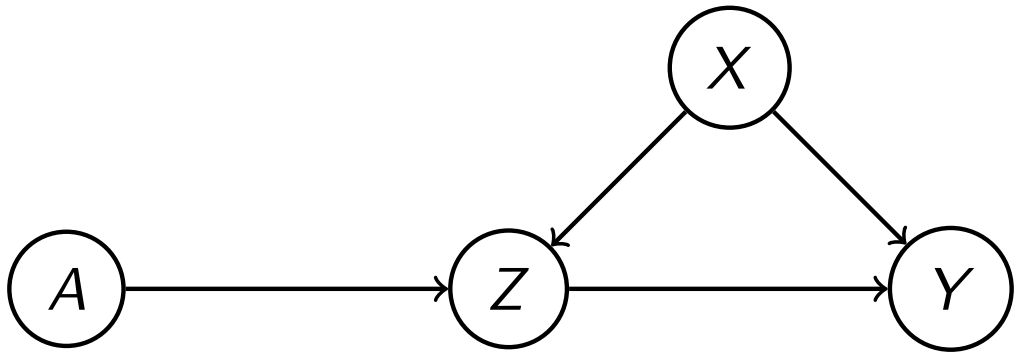
이를 DAG로 그리면, X와 U가 중복되지 않는다는 가정 하에,



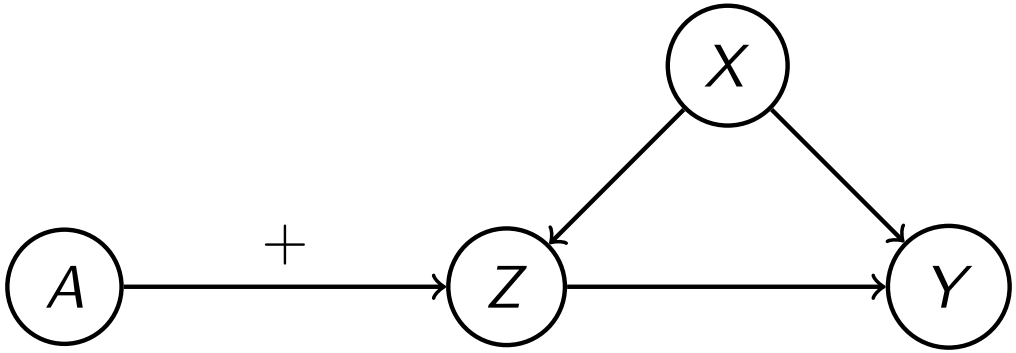
Instruments or Encouragement



우리는 디자인 과정에서, $U = A$를 '금연 권고'로 하려 합니다. 이처럼 도구변수 자체가 Z를 cause한다면, 이를 **causal instrument**라 합니다(원래 상황인 경우, A는 **surrogate instrument**). 이는 아래처럼 요약됩니다.

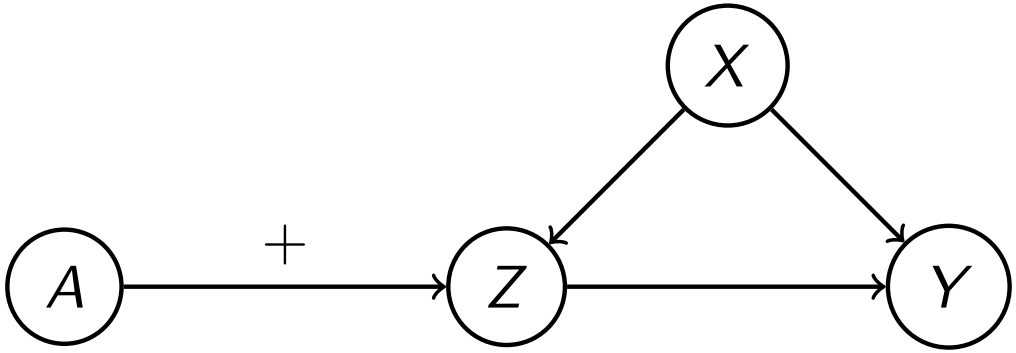


Instruments or Encouragement



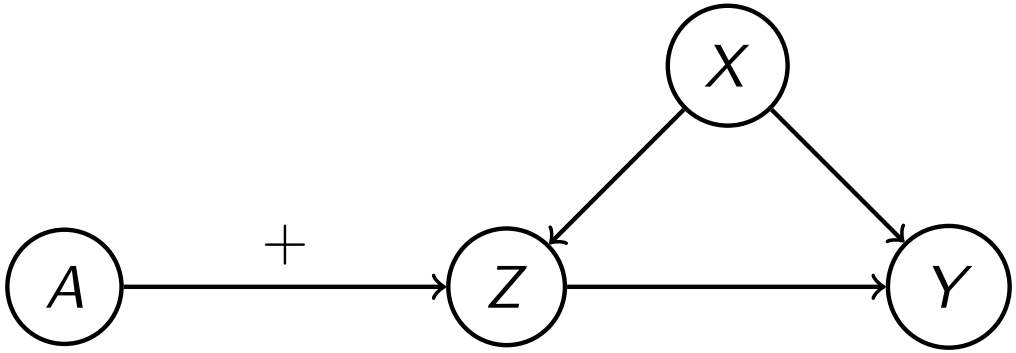
금연 권고는 직접적으로 금연을 촉진합니다. 따라서 $A \to Z$는 양의 효과를 갖는 인과효과를 묘사합니다. 동시에, 도구변수의 3번 조건으로부터 A와 Y는 동일한 원인을 공유하지 않아야 합니다. 가장 간단하게는 금연 권유를 랜덤하게 수행하는 경우를 생각해볼 수 있습니다. $A \to Y$의 직접적인 경로는 없고, A는 오직 $A \to Z \to Y$를 통해서만 Y에 영향을 미치게 됩니다. 이제 감이 오시나요?

Instruments or Encouragement



DAG로부터 A와 Y의 상관관계는 오직 이 인과관계 경로로부터 나오게 됨을 압니다. 이는 내생성이 없는 것으로부터도 알 수 있습니다. 또한 A와 Z의 상관관계 역시 $A \to Z$의 인과관계로부터만 나옵니다. 그렇다면 $A \to Z \to Y$ 경로의 효과와 $A \to Z$ 경로의 효과를 회귀분석으로 추정할 수 있기에, 그 효과를 분리하면 $Z \to Y$라는 직접적인 인과관계 경로의 효과 역시 identify 할 수 있게 되는 것입니다.

Instruments or Encouragement



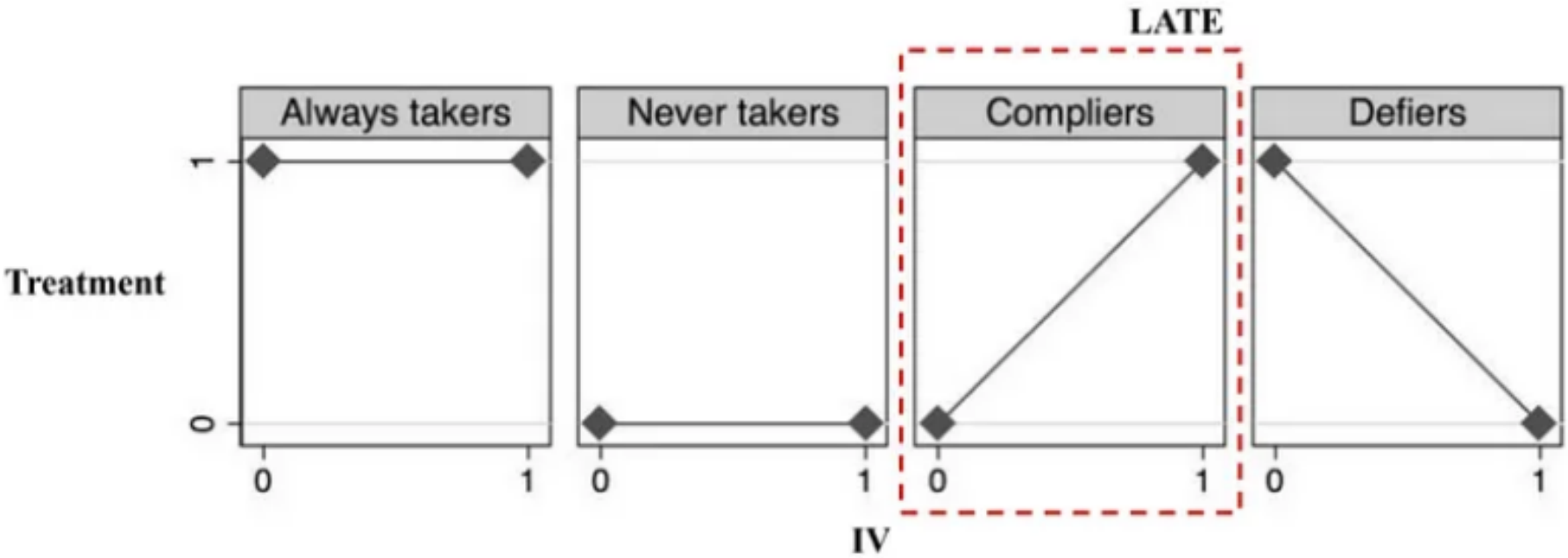
심리학, 의학 등의 연구에서는 이러한 A를 도구변수라는 이름 대신 encouragement라고 부르기도 합니다. 연구자의 제안, 혹은 유도를 통하여 이루어지는 경우가 많기 때문입니다. 반면 경제학과 같은 사회과학에서는 그러한 상황이 많이 발생하지는 않기 때문에, 도구변수의 세 조건을 만족하는 A를 미리 수집하고 이를 identify를 위한 도구로써 사용한다는 점에서 도구변수라고 부릅니다. 또한 많은 경우 이는 해당 학문의 틀과 모형 하에서 미리 결정된 후 데이터 수집 단계에서 이를 추가적으로 조사합니다.

Instruments or Encouragement

먼저 A가 binary인 경우를 고려해 보겠습니다. 이제 potential outcome framework를 이용하여, 각 i의 Z_i 역시도 A_i에 따른 결과로 볼 수 있습니다. 즉 $Z_i(1)$과 $Z_i(0)$이 존재하며, $Z_i(A_i)$만이 관측가능하고 $Z_i(1-A_i)$는 관측 불가한 것입니다. 그렇다면 각 개체는 네 종류로 구분할 수 있습니다.

1. **Never-taker**: $Z_i(1) = Z_i(0) = 0$, 권고를 받든 말든 항상 Z를 수행하는 사람입니다.
2. **Always-taker**: $Z_i(1) = Z_i(0) = 1$, 권고를 받든 말든 항상 Z를 수행하는 사람입니다.
3. **Complier**: $Z_i(1) = 1, Z_i(0) = 0$, 권고를 받으면 Z를 수행하고, 그렇지 않으면 수행하지 않는 사람입니다.
4. **Defier**: $Z_i(1) = 0, Z_i(0) = 1$, 청개구리같이 권고를 받으면 Z를 수행하지 않고, 그렇지 않으면 수행하는 사람입니다.

Instruments or Encouragement



Monotone Treatment Selection(MTS): #Compliers > #Defiers
i.e. $P(Z = 1|A = 1) > P(Z = 1|A = 0)$

Instruments or Encouragement

Monotonicity Assumption: no defiers. i.e. $Z_i(1) \geq Z_i(0)$
(Note that it is stronger than MTS.)

이때 도구변수에 의하여 처리 여부가 변화하여, 처리의 인과적 효과를 추정할 수 있는 집단은 complier 집단 뿐입니다. 이에 따라서, **LATE**(Local Average Treatment Effect), 혹은 CATE(Complier Average Causal Effect)는 아래처럼 정의합니다.

$$\begin{aligned} LATE &= \mathbb{E}[Y(1) - Y(0) | Z(1) = 1, Z(0) = 0] \\ &= \mathbb{E}[Y|A = 1] - \mathbb{E}[Y|A = 0] \end{aligned}$$

즉 이는 complier들에 local하게 작용하는 ATE가 됩니다.

Instruments or Encouragement

이제 이 LATE를 어떻게 identify하는지 확인하여 보겠습니다. 먼저,

$$\begin{aligned}
&\mathbb{E}[Y|A=1] - \mathbb{E}[Y|A=0] \\
&= \sum_{\text{group}} (\mathbb{E}[Y|A=1, \text{group}] - \mathbb{E}[Y|A=0, \text{group}]) P(\text{group}) \\
&= (\mathbb{E}[Y|A=1] - \mathbb{E}[Y|A=0]) P(\text{complier})
\end{aligned}$$

를 얻습니다. 이때 group $\in \{$never-taker, always-taker, complier$\}$입니다.

- $P(\text{group}|A=1) = P(\text{group}|A=0) = P(\text{group})$
  ($A \to Z$ 경로는 $Z(A)$만을 통하고, 내생성 문제가 없기에)
- never-taker, always-taker에 대하여, $\mathbb{E}[Y|A, \text{group}]$은 A에 무관
- complier에 대하여, $\mathbb{E}[Y|A, \text{group}]$은 A에만 의존(Why?)

Instruments or Encouragement

그리고 Monotonicity Assumption이 MTS를 내포하므로,

$$\begin{aligned} 0 &< \mathbb{E}[Z|A=1] - \mathbb{E}[Z|A=0] \\ &= P(Z=1|A=1) - P(Z=1|A=0) \\ &= P(\text{always-taker}) + P(\text{complier}) - P(\text{always-taker}) \\ &= P(\text{complier}) \end{aligned}$$

이 성립하게 됩니다. 따라서 IV method의 **usual IV estimand**는

$$\frac{\mathbb{E}[Y|A=1] - \mathbb{E}[Y|A=0]}{\mathbb{E}[Z|A=1] - \mathbb{E}[Z|A=0]}$$

로 identify되며, 이를 통해 얻은 추정량을 **Wald estimator**라 부릅니다.

Instruments or Encouragement

$\mathbb{E}[Y|A=a], \mathbb{E}[Z|A=a]$는 단순히 sample average

$$\mathbb{E}_n[Y|A=a], \quad \mathbb{E}_n[Z|A=a]$$

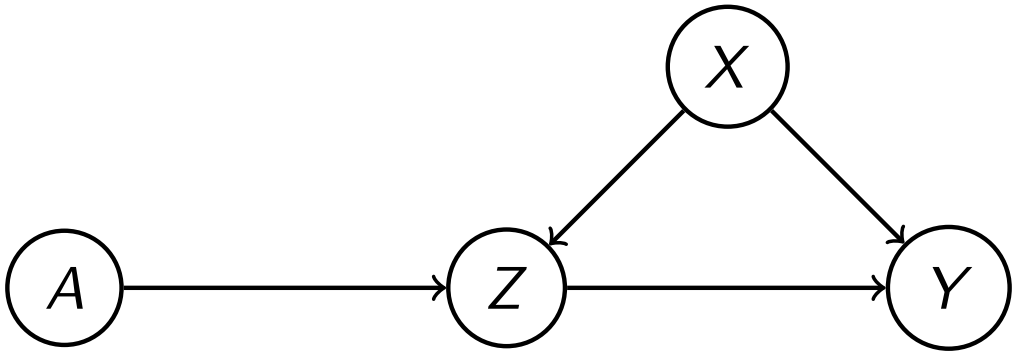
를 사용해 추정해주면 됩니다. 따라서 Wald estimator는

$$\widehat{LATE} = \frac{\frac{1}{n_{A=1}}\sum_{i=1}^n A_i Y_i - \frac{1}{n_{A=0}}\sum_{i=1}^n (1-A_i) Y_i}{\frac{1}{n_{A=1}}\sum_{i=1}^n A_i Z_i - \frac{1}{n_{A=0}}\sum_{i=1}^n (1-A_i) Z_i}$$

으로 주어집니다. 이는 A가 causal instrument가 아니라 surrogate instrument 여도 여전히 성립합니다.

2SLS

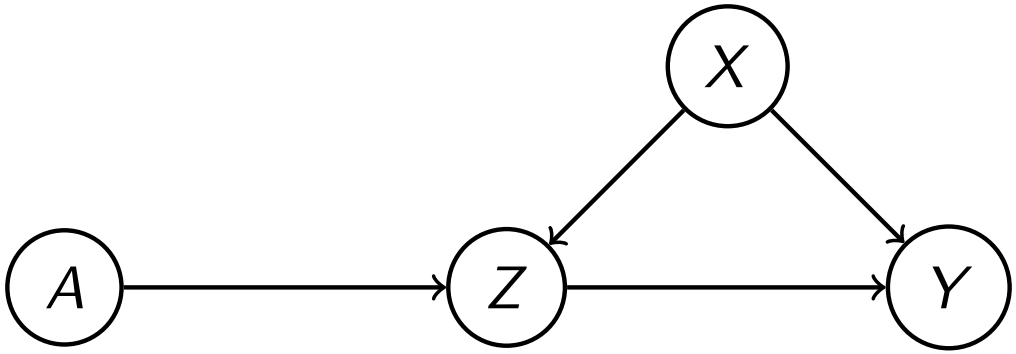
그렇다면 이를 더 확장하여서, Z나 A가 binary가 아니어도 되면 어떨까요? 단변량이 아니라 다변량이라면 어떨까요? 이 경우에도 Wald estimator를 확장해줄 수 있습니다. 다시 아래 모형을 생각해 볼까요?



확률변수가 아니라 결정적(deterministic) 함수라고 생각해 봅시다. 그러면

$$y = f(x, z)$$
$$z = g(a, x)$$

처럼 생각할 수 있습니다.

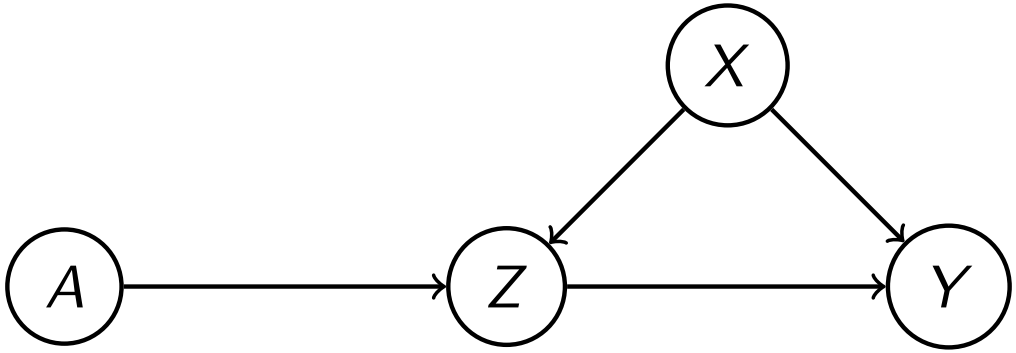


이제 $Z \to Y$ 경로의 효과는

$$\frac{\partial y}{\partial z} = f_z(x, z)$$

으로 주어집니다. 그런데 z는 $g(a, x)$로 표현되므로,

$$\frac{\partial y}{\partial a} = \frac{\partial y}{\partial z} \cdot \frac{\partial z}{\partial a}$$



따라서 단변량의 경우

$$\frac{\partial y}{\partial z} = \frac{\partial y/\partial a}{\partial z/\partial a}$$

입니다. 어디서 많이 보지 않았나요? 바로 Wald estimator의 형태와 동일합니다. 분모에는 a에 따른 z의 인과적 변화, 분자에는 a에 따른 y의 인과적 형태가 들어가게 됩니다.

2SLS

벡터의 경우에는 $\partial y/\partial z$가 $\partial y/\partial a$를 반응변수로, $\partial z/\partial a$를 설명변수로 하여 회귀분석을 적합하는 상황과 동일합니다. f와 g가 선형함수라고 해 보겠습니다. 그렇다면

$$Z = A\gamma + X\delta + \eta$$
$$Y = Z\beta + X\alpha + \epsilon$$

로 모형화할 수 있습니다. 이때 $A \perp\!\!\!\perp X$이며, ϵ과 η는 완전히 무작위적인 변동으로 모든 변수에 독립이라고 할 수 있습니다. 그렇다면, 이를 X와 A에 대한 식으로 요약하면

$$Y = A(\gamma\beta) + X(\alpha + \beta\delta) + (\eta\beta + \epsilon)$$

이고 오차항을 $u = X(\alpha + \beta\delta) + (\eta\beta + \epsilon)$으로 쓰면 $A \perp\!\!\!\perp u$입니다. 동일하게 $v = X\delta + \eta$ 역시 $A \perp\!\!\!\perp v$입니다.

2SLS

그렇다면 Y, Z, A에 대한 모형

$$Z = A(\gamma) + v \qquad \text{(first stage)}$$
$$Y = A(\gamma\beta) + u \qquad \text{(second stage)}$$

는 내생성에서 자유롭습니다. 그렇다면

$$\widehat{\gamma\beta} = (A^T A)^{-1} A^T Y$$
$$\hat{\gamma} = (A^T A)^{-1} A^T Z$$

일 것이고, 둘의 적절한 비를 구하면 β의 추정량을 구할 수 있게 됩니다. A, Z가 1차원인 경우에는 그 둘의 비가 단순히 Wald estimator로 주어집니다. 더욱 일반적인 벡터 상황에서는 어떨까요?

2SLS

2SLS(2 Stage Least Squares) Estimator는 두 단계를 통해 구해집니다.

1. 먼저, Z를 반응변수로, A를 설명변수로 하는 모형 $Z = A\gamma + v$를 적합하여 A를 이용한 Z의 추정량 $\tilde{Z} = A\hat{\gamma} = A(A^TA)^{-1}A^TZ$를 구한다.
2. 그 다음, Y를 설명변수로, $\tilde{Z}$를 반응변수로 하는 모형 $Y = \tilde{Z}\beta + u$를 적합하며 $\hat{\beta}^{2SLS}$를 구한다. 이는 앞선 페이지에서 $Y = (A\gamma)\beta + u$를 적합하는 상황에서 $A\gamma$ 대신 $A\hat{\gamma}$를 쓴 것과 같다.

2SLS estimator는 간단하게 아래처럼 쓸 수 있습니다.

$$\hat{\beta}^{2SLS} = (\tilde{Z}^T\tilde{Z})^{-1}\tilde{Z}^TY = (Z^TA(A^TA)^{-1}A^TZ)^{-1}Z^TA(A^TA)^{-1}A^TY$$

2SLS

$$\hat{\beta}^{2SLS} = (\tilde{Z}^T\tilde{Z})^{-1}\tilde{Z}^T Y = (Z^T A(A^T A)^{-1}A^T Z)^{-1} Z^T A(A^T A)^{-1}A^T Y$$

2SLS estimator는 아래와 같은 성질을 가집니다.

- 도구변수가 적절히 선택된 경우, 비편향추정량이다. 그러나 많은 경우 편향이 있으며, 무의미한 도구변수가 증가할수록 편향이 증가하는 성질이 있다.
- 그럼에도, 일치추정량이다.
- 표본의 개수가 커질 때, $\hat{\beta}^{2SLS} \dot{\sim} N(\beta, (X^T A(A^T A)^{-1}A^T X)^{-1})$
- 일반적으로 OLS 추정량에 비해 그 분산은 크다. 즉 $\text{Var}(\hat{\beta}^{OLS}) \leq \text{Var}(\hat{\beta}^{2SLS})$.

Python Practice

Angrist, J. D., & Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings?. The Quarterly Journal of Economics, 106(4), 979-1014.

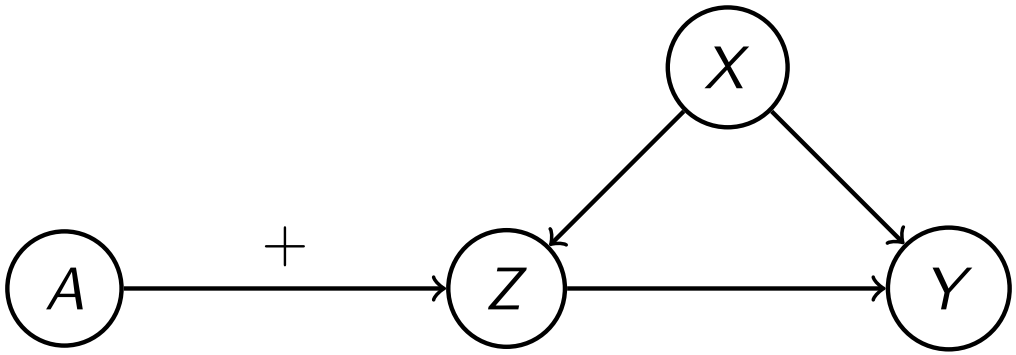
Research Question: 교육이 임금에 미치는 영향은 어떠한가?

미국의 compulsory school attendance: 아이들이 학교에 입학하려면 그 해 1월 1일에 6살이 되어야 하고, 16살이 되어서야 중퇴할 수 있다. 따라서 연초에 태어난 아이들은 더 많은 나이에 학교에 입학하게 되며, 더 빨리 중퇴할 수 있다. 따라서 늦게 태어날수록 더 많은 교육을 받게 된다.

Quiz: 이 연구에서 Z는 교육 수준, Y는 임금이다. A는?

Python Practice

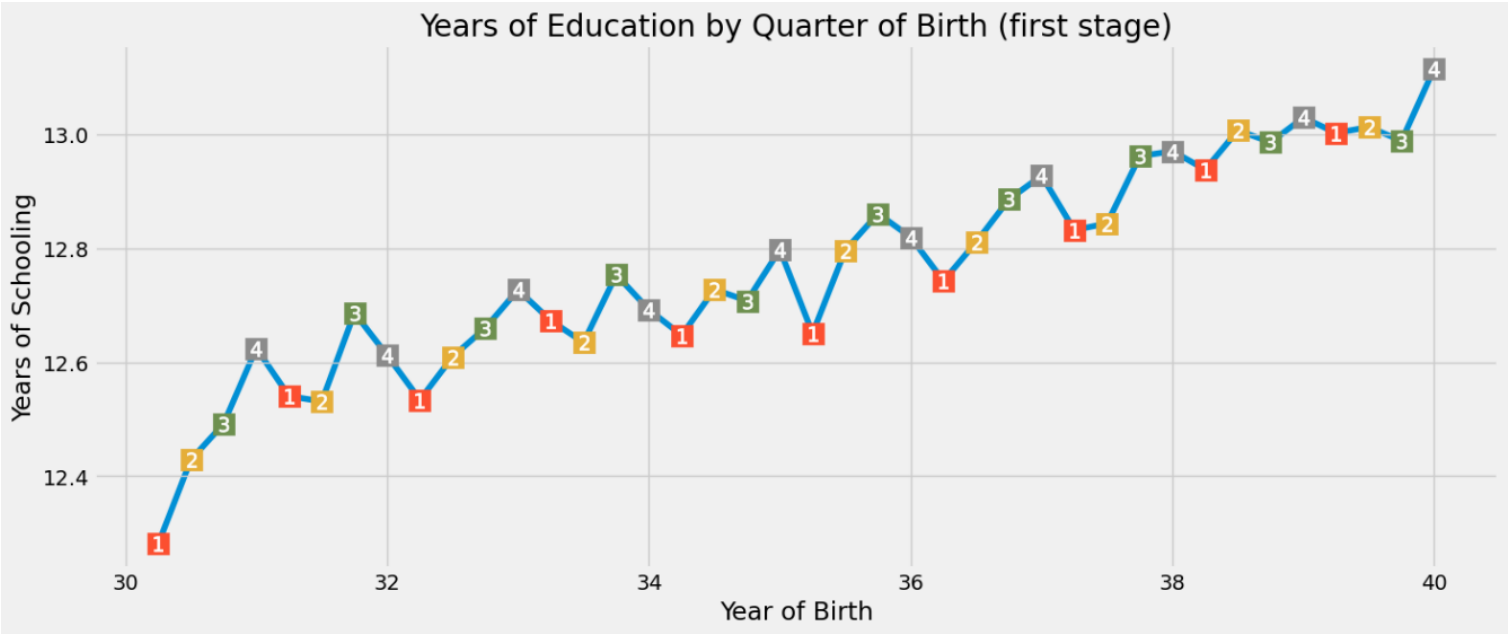
우리는 A를 분기, 특히 4분기 출생 여부로 하기로 합시다. 그렇다면



처럼 causal diagram을 그릴 수 있습니다. 출생한 분기는 교육 기간에는 영향을 강하게 미칠 것이지만, 임금에는 그리 큰 직접적 영향을 미치지 않을 것입니다.

Python Practice

먼저, Wald Estimator를 구해 보겠습니다. 이는 Y를 A에 회귀분석하여 얻은 값을 Z를 A에 회귀분석하여 얻은 값으로 나누면 됩니다. 이를 위하여, A가 Z에 미치는 영향을 먼저 보겠습니다.



Python Practice

회귀 결과는 아래와 같습니다.

Covariance Type:		nonrobust				
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	12.7473	0.007	1937.396	0.000	12.734	12.760
q4	0.0921	0.013	6.935	0.000	0.066	0.118

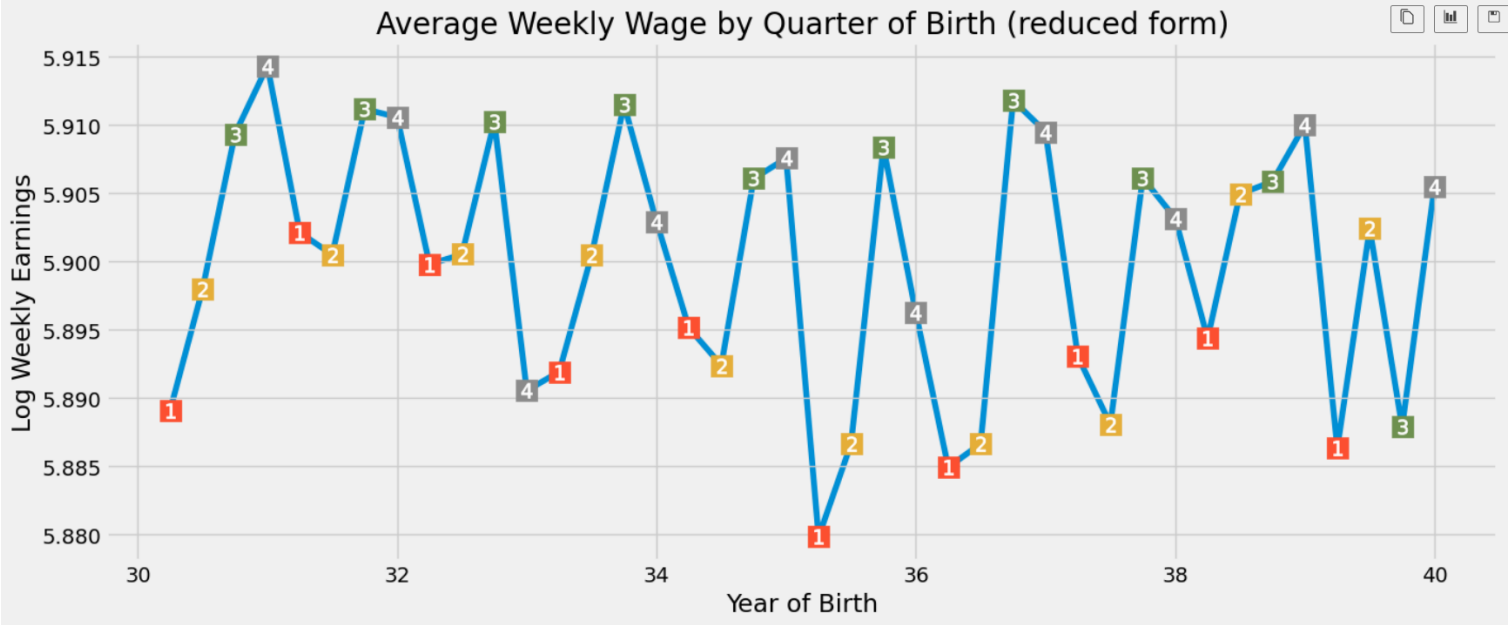
즉 이를 다르게 말하면,

$$\mathbb{E}[Z|A=1] - \mathbb{E}[Z|A=1] = 0.0921$$

이 되겠네요.

Python Practice

다음으로 A가 Y에 미치는 영향을 보겠습니다.



Python Practice

회귀 결과는 아래와 같습니다.

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	5.8983	0.001	4332.898	0.000	5.896	5.901
q4	0.0068	0.003	2.479	0.013	0.001	0.012

즉 이를 다르게 말하면,

$$\mathbb{E}[Y|A=1] - \mathbb{E}[Y|A=1] = 0.0068$$

이 되겠습니다. 따라서 Wald 추정량은 아래처럼 계산됩니다.

$$\widehat{LATE} = \frac{\mathbb{E}[Y|A=1] - \mathbb{E}[Y|A=1]}{\mathbb{E}[Z|A=1] - \mathbb{E}[Z|A=1]} = 0.074$$

Python Practice

혹은 Wald 추정량을 사용하지 않고 직접 두 단계의 회귀분석을 돌려 그 결과를 얻을 수도 있습니다. 그 경우에도 마찬가지로, 동일한 추정량을 얻습니다 (코드 2절 참고).

가장 간단하게는 `linearmodels.iv.model` 모듈의 함수 `IV2SLS()`를 이용할 수도 있습니다. 여기에서 변수는 네 가지로 구성됩니다.

- `dependent`: Y 역할을 할 반응변수
- `exog`: 외생변수. 지금 우리는 상수항을 여기서 추가합니다.
- `endog`: Z 역할을 할 처리변수, 혹은 설명변수
- `instruments`: A 역할을 할 도구변수

여기에서도 마찬가지로 동일한 추정량을 얻습니다(코드 3절 참고).

Python Practice

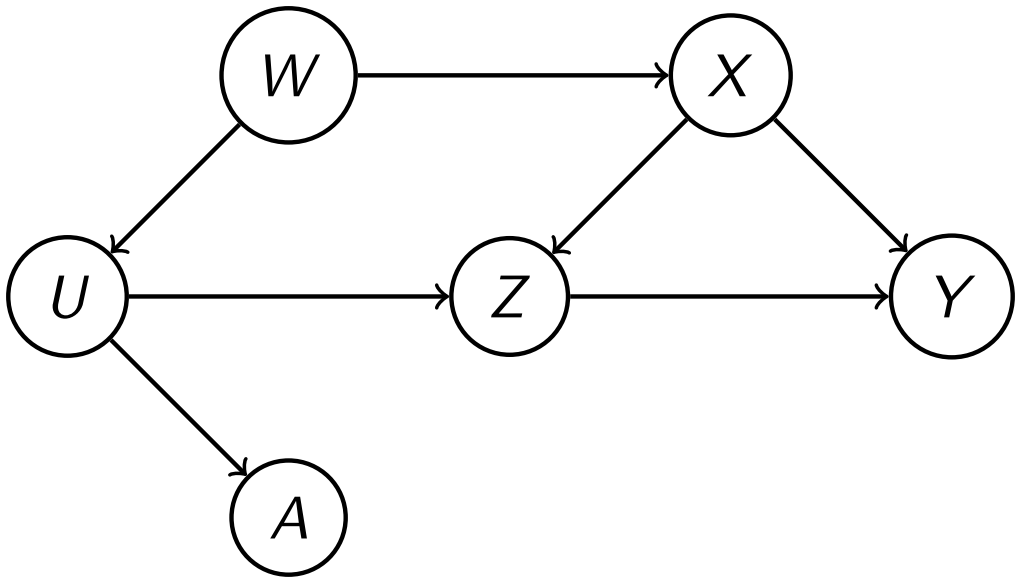
Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
constant	4.9555	0.3577	13.852	0.0000	4.2543	5.6566
years_of_schooling	0.0740	0.0280	2.6401	0.0083	0.0191	0.1289

함수를 이용하면 쉽게 표준오차, t통계량, p값, 신뢰구간 등을 얻을 수 있습니다. 우리의 경우 교육시간의 증가가 임금의 증가로 이어진다는 결과가 통계적으로 유의함을 확인해줄 수 있겠습니다.

Question 1. Exogeneous Variables

만약 우리가 추가적인 자료 W들을 모을 수 있다면, 이는 X와 U에 대한 정보를 추가적으로 제공함으로써 2SLS 과정에서 추정량의 분산을 줄일 수 있을 것입니다. 이에 따라 2SLS의 회귀모형에 외생변수를 넣는 상황을 고려할 수 있습니다.



Question 1. Exogeneous Variables
더 나아가서는 사실 W가 X와 U를 직접적으로 cause하지는 않더라도, 적절한 상관관계를 가져 2SLS 과정에서 X와 U에 의한 변동을 설명할 수 있다면 이는 추정량의 분산 감소에 영향을 미칠 수 있습니다. 그렇다면

$$Z = A\gamma + W\xi + v \qquad \text{(first stage)}$$
$$Y = \tilde{Z}\beta + W\zeta + u \qquad \text{(second stage)}$$

에서 얻은 $\hat{\beta}$가 새로운 2SLS 추정량입니다.

우리는 exog 변수에 W를 추가함으로써 쉽게 이를 수행해줄 수 있습니다. 이러한 W를 **공변량**(covariates)이라 부르기도 합니다.

Question 2. Verifying Instruments

Instrument가 될 세 가지 조건을 다시 생각해 보겠습니다.

(i) A는 Z와 상관관계를 가진다.

(ii) A는 Y에 직접적인 인과효과를 가지지 않는다. 즉 A의 Y에의 효과는 Z를 거치는 간접효과만 있다.

(iii) A와 Y는 동일한 원인을 공유하지 않는다.

이 중 (ii)와 (iii)은 직접적으로 확인할 수 없습니다. 즉 이는 사실인 것으로 믿거나, 모형 등을 통해 밝혀야만 합니다. 우리가 검정할 수 있는 것은 사실 (i) 뿐입니다. 이는 간단합니다.

$$Z = A\gamma + \epsilon$$

을 적합한 뒤 γ가 통계적으로 유의하게 0이 아닌지 확인하면 됩니다.

Question 3. Weak Instrument

도구변수가 약하다, 혹은 **weak instrument**라는 것은 아래처럼 두 방법으로 정의됩니다.

- (substantively weak): 만약 $A - Z$의 실제 상관관계가 매우 빈약한 경우
- (statistically weak): 만약 $A - Z$의 추정된 상관관계가 빈약한 경우, 즉 F통계량이 10보다 작은 경우.

이들 모두는 왈드 추정량에서 분모 부분을 작게 만들어 전반적인 추정량의 분산을 높인다는 약점을 가지게 합니다. 이는 추정량의 신뢰구간을 넓히므로, 유효한 추정을 어렵게 만듭니다. 이러한 weak instrument 하에서도 추론을 할 수 있는 방법 역시 많이 개발되어 있습니다.

Question 4. Checking Endogeneity
한편 우리는 사실 DAG를 예측하기만 할 뿐 실제 관계는 알 수 없으므로, 모형에 근거한 2SLS보다 그냥 OLS가 낫지 않은지 고민이 될 수 있습니다. 사실 우리가 가장 원하는 것은 표본이 증가하면 이들 값이 실제 값으로 다가간다는 믿음, 즉 consistency이므로, 두 추정량의 consistency를 비교하는 것이 두 추정량 중 어떤 것이 더 나은지 이야기해 줍니다. 더 나아가서는, 실제로 내생성이 있는지 여부까지 대략적으로 생각해볼 수 있습니다. (Why?)

일반적으로, **Durbin-Wu-Hausman test**를 이용해 내생성을 판단합니다. 그 검정통계량은 아래와 같습니다.

$$H := (\hat{\beta}^{OLS} - \hat{\beta}^{2SLS})^T (\text{Var}(\hat{\beta}^{2SLS}) - \text{Var}(\hat{\beta}^{OLS})^{-1} (\hat{\beta}^{OLS} - \hat{\beta}^{2SLS})$$

이는 점근적으로 카이제곱분포를 따름이 알려져 있습니다.

Question 4. Checking Endogeneity

아주 간단하게, `wu_hausman()` 함수를 이용하면 이를 검정할 수 있습니다.

```
iv_model_cov.wu_hausman()
✓ 12.7s

Wu-Hausman test of exogeneity
H0: All endogenous variables are exogenous
Statistic: 4.4688
P-value: 0.0345
Distributed: F(1,329447)
WaldTestStatistic, id: 0x23124984590
```

여기에서 귀무가설이 기각되었으므로, 내생성이 있음을 대략적으로 알 수 있습니다. 이외에도 Wooldride's regression test 등이 이용될 수 있습니다.

Question 5. Covariance of the 2SLS Estimator

IV-2SLS 추정량은 등분산 가정에 근거하고 있습니다. 그러나 OLS 상황과 같이, 여기에서도 이분산강건 표준오차의 계산과 이를 통한 추론이 가능합니다. `fit()` 함수를 이용할 때, `cov_type`을 `robust`로 지정하면 됩니다. 추가적으로 `debiased` 옵션 역시 covariance estimator를 조정하는 역할을 합니다. 이 경우에도 그 추정량의 값은 달라지지 않으나, 표준편차가 달라져 유의성 검정 등의 결과는 바뀔 수 있습니다.

$$\widehat{\text{Var}}_{HC}(\hat{\beta}^{2SLS}) = (\tilde{Z}^T\tilde{Z})^{-1}(\tilde{Z}^T\sum_{i=1}^{n}\hat{u}_i^2\tilde{Z})(\tilde{Z}^T\tilde{Z})^{-1}$$

전반적으로 OLS와 유사하나, Z 대신 $\tilde{Z}$가 사용된다는 점만 다르다고 볼 수 있겠습니다.

Question 6. Number of Instruments and IV-GMM

Regression Discontinuity Design(RDD)

RDD

Take-Home Messages

Next Session

주제: Nonparametric Inference and Statistical Causal Inference

- Nonparametric Inference
  - Parametric Model and Semi/Non-parametric Model
  - Wilcoxon Rank-Sum Test and Huber's M-estimator
  - Bootstrap
  - Kernel Density Estimation and Nonparametric Regression
- Matching
  - Propensity Score and Unconfoundedness
  - Propensity Score Matching(PSM) and Other Matching Designs
- Weighting
  - Inverse Probability Weighting
  - Doubly Robust Estimator
- Causal Inference in Panel Data
  - Difference-in-Differences Design and Two-way Fixed Effects Model
  - Synthetic Control Method