

<Simulation code 해야 하는 것 정리>

Data generation

$$X = [X_B, X_C, X_U] \sim N(0, I) \quad / \quad B = \begin{cases} 1 & (X_B \geq 0) \\ 0 & (X_B < 0) \end{cases} \quad C = X_C, \quad U = P(X_U < \alpha) \rightarrow U \sim \text{unif}(0,1)$$

$B \sim \text{Ber}(0.5)$
 $C \sim N(0,1^2)$

$E(\text{exposure 할당}) \sim \text{Ber}(p_Z)$ 이때, $p_Z = \text{logit}^{-1}(\beta_0 + \beta_B B + \beta_C C)$ random 추출할 때 set.seed 사용!

$\Rightarrow \beta_0, \beta_B, \beta_C$ 가 결정되면, $E(\text{exposure})$ 변수 생성 가능! 이때, $\beta_0, \beta_B, \beta_C (0,0,0) \sim (1,1,1)$ 까지 변화시키며

각 $(\beta_0, \beta_B, \beta_C)$ 당 1000개의 sample 만들어 $\hat{P}(E=1) = \frac{1}{1000} \sum_{i=1}^{1000} E_i$ 추정해보자!

$\Rightarrow \beta_B = \beta_C = 0.5$ 로 고정, β_0 바꾸어가며 $\hat{P}(E=1)$ 값 낮춰보자! / 횟수를 100,000으로 늘려!

Next) $P_{Y|U,B,C,E}$ 결정!
 $\begin{cases} Y \text{가 연속형}; Y = \sigma^T(u, B, C, E) + \epsilon \\ Y \text{가 이산형}; P(Y=1|u, B, C, E) = \text{logit}^{-1}(\sigma^T(u, B, C, E)) \end{cases}$
 $\{ \log 1.2, \log 1.5, \log 2, \log 2 \}$ 로 고정

$\Rightarrow \beta_0, \beta_B, \beta_C$ 1개씩 지정하면, 우리는 $(\underbrace{B, C, U, E}_X, Y) \sim P$ 하나 가지게 된다.

Y가 연속형일 때만 해보기!

Next) $(X_T, E_T, Y_T) \sim P^{\text{IID}} \quad (T=1, \dots, 100) \Rightarrow$ Data 이용해 $\hat{ATE}^{\text{IPW}}, \hat{ATE}^{\text{matching}}, \hat{ATE}^{\text{regression}}$

구해보자! 여기서 U는 포함되지 않는다.

\hat{ATE}^{IPW} 도 계산해보기!

\Rightarrow PSM (1:1 matching)

공변량간 거리 이용해 greedy 1:1 matching

Next) 추정한 값과 실제 effect 비교해 성능 평가! ($\Rightarrow Y$ 가 연속형일 때, $\hat{ATE} = \hat{ATT} = \sigma^T$ 의 4번째 원소 $= \log 2$)

\Rightarrow 원래는 X의 함수 형태!

cf) code 최대한 간단하게 작성!