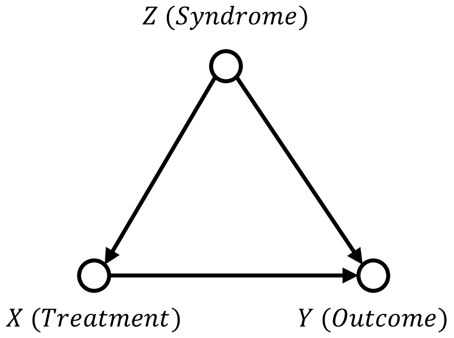
Study question 3.2.1 문제에서 Study question 1.5.2를 참고하기에 다시 가져옴.

문제 정리

· $P(Z=1) = r$

X	Y	Z	P
1	1	0	$P(Y=1 \mid Z=0, X=1) = p_2$
0	1	0	$P(Y=1 \mid Z=0, X=0) = p_1$
0	1	1	$P(Y=1 \mid Z=1, X=0) = p_3$
1	1	1	$P(Y=1 \mid Z=1, X=1) = p_4$

X	Z	P
1	0	$P(X=1 \mid Z=0) = q_1$
1	1	$P(X=1 \mid Z=1) = q_2$



중요한 것은 graphical model을 기반으로 한다는 것이다.

· Z는 어떠한 node의 child가 되지 않기에 그 자체로 확률이 존재한다.

· X는 Z의 child 이기 때문에 확률을 구할 때 Z를 고려한다.

· Y는 X와 Z 둘 모두의 child 이기에 이 둘을 모두 고려한다.

Study question 3.2.1 (a), (b)

a. Graph의 구조가 동일하기 때문에 adjustment formula를 적용하면 쉽게 풀 수 있다.
(당연하게도 Solution에 깔끔하게 정리되어 있다)

· $P(Y=0|do(X=0))$

$=P(Y=0|X=0,Z=0)P(Z=0)+P(Y=0|X=0,Z=1)P(Z=1)$

$=(1-p_1)(1-r)+(1-p_3)r$

$=1-r-p_1+rp_1+r-rp_3$

$=1-(rp_3+(1-r)p_1)$

· $P(Y=1|do(X=0))$

$=P(Y=1|X=0,Z=0)P(Z=0)+P(Y=1|X=0,Z=1)P(Z=1)$

$=p_1(1-r)+p_3r$

· $P(Y=0|do(X=1))$

$=P(Y=0|X=1,Z=0)P(Z=0)+P(Y=0|X=1,Z=1)P(Z=1)$

$=1-(rp_4+(1-r)p_2)$

· $P(Y=1|do(X=1))$

$=P(Y=1|X=1,Z=0)P(Z=0)+P(Y=1|X=1,Z=1)P(Z=1)$

$=rp_4+(1-r)p_2$

각 합이 1
이 된다.

b. a 또한 adjustment formula를 통해 계산하였기에 a와 동일하다.

Study question 3.2.1 (c)

c.

· $ACE = P(y_1|do(x_1)) - P(y_1|do(x_0))$

$= \{rp_4 + (1-r)p_2\} - \{p_1(1-r) + p_3 r\}$

· $RD = P(y_1|x_1) - P(y_1|x_0)$

$= \frac{P(y_1, x_1)}{P(x_1)} - \frac{P(y_1, x_0)}{P(x_0)}$

$= \frac{[rp_4 q_2 + (1-r)p_2 q_1]}{[rq_2 + (1-r)q_1]} - \frac{[rp_3(1-q_2) + (1-r)p_1(1-q_1)]}{[r(1-q_2) + (1-r)(1-q_1)]}$

· $ACE - RD$가 최소가 되는 경우

① $r = 0, q_1 \neq 0, q_1 \neq 1$

② $r = 1, q_2 \neq 0, q_2 \neq 1$

· 나는 다음과 같이 계산하였었다.

$RD = (p_2 + p_4) - (p_1 + p_3)$

→ $P(Y = y|X = x, Z = z)$를 그대로 가져다 쓴 경우인데 지금의 예는 Z가 조건으로 주어지지 않았기에 위처럼 계산하면 안 된다.

· $\frac{P(y_1, x_1)}{P(x_1)} - \frac{P(y_1, x_0)}{P(x_0)}$이 부분을 구할 때는 다음을 이용한다.

(*Product decomposition* 이용)

① p29의 Rule of product decomposition을 이용

예) $P(y_1, x_1) = P(z_0, y_1, x_1) + P(z_1, y_1, x_1)$

$P(y_1|x_1, z_0)$는 Y의 parents가 X와 Z이기 때문이다.

$P(z_0)P(x_1|z_0)P(y_1|x_1, z_0)$

② 기존 식을 변환

예) $P(x_1) = P(z_0, x_1) + P(z_1, x_1)$

$P(z_0)P(x_1|z_0)$

Conditional probability가 아니라 Joint probability를 구한다는 것을 매우 조심해야 한다.

· *Solution*에는 ①만 나와있지만 ②도 가능하다

Why? 왜 r을 기준으로 성립하는가?

→ $P(Z = 1) = r = 0$ *or* 1

→ 이 경우 Z값은 단 하나가 된다
($Z = 0$이 0% 혹은 100%가 되기 때문)

→ 결과적으로 Y값은 오로지
X값에 의해서만 영향을 받는다.

→ 그렇기에 intervention의 의미가 없어진다?

Study question 3.2.1 (d)

d. · 다음을 보이면 된다.

① $The\ sign\ of\ P(y|x_1, z) - P(y|x_0, z) = The\ sign\ of\ ACE$ (For all value z of Z)

→ 1.5.2 (c)를 통해 주어진 parameter는 $Simpson's\ paradox$를 만족하는 것은 확인 되었다.

② ①의 부호와 전체 결과($P(y_1|x_1) - P(y_1|x_0)$)의 부호가 반대이면 이번 문제에서 $Simpson's\ paradox$가 성립함을 보일 수 있다.

→ $P(y_1|x_1) - P(y_1|x_0) = RD$이기에 (c)를 이용한다.

→ $\frac{[rp_4q_2+(1-r)p_2q_1]}{[rq_2+(1-r)q_1]} - \frac{[rp_3(1-q_2)+(1-r)p_1(1-q_1)]}{[r(1-q_2)+(1-r)(1-q_1)]}$, 에 Solution에 주어진 parameters $[p_1 = 0.1, p_2 = 0, p_3 = 0.3, p_4 = 0.2, q_1 = 0, q_2 = 1, r = 0.1]$를 통해 직접 계산을 하여 부호를 확인한다.

→ $\frac{[rp_4q_2+\cancel{(1-r)p_2q_1}]}{[rq_2+\cancel{(1-r)q_1}]} - \frac{[\cancel{rp_3(1-q_2)}+(1-r)p_1(1-q_1)]}{[\cancel{r(1-q_2)}+(1-r)(1-q_1)]}$

→ $\frac{rp_4q_2}{rq_2} - \frac{(1-r)p_1(1-q_1)}{(1-r)(1-q_1)}$

→ $p_4 - p_1$

→ $0.2 - 0.1 = 0.1$

→ $Solution$에서 ACE는 -0.1이기에 부호가 반대이다.
즉, $Simpson's\ paradox$ 성립.

Study question 3.3.1

a. Backdoor paths를 구하고, backdoor criterion을 만족하는 변수집합을 구한다.

· *Backdoor paths*

a. X←A←B→Z→Y

b. X←Z→Y

c. X←Z←C→D→Y

d. X←A←B→Z←C→D→Y (Condition on Z, Collider)

· Backdoor criterion을 만족하는 set of variables

a. $Sets\ of\ 2\ nodes$: $\{Z,A\},\{Z,B\},\{Z,C\},\{Z,D\}$

b. $Sets\ of\ 3\ nodes$: $\{Z,A,B\},\{Z,A,C\},\{Z,A,D\},\{Z,B,C\},\{Z,B,D\},\{Z,C,D\}$

c. $Sets\ of\ 4\ nodes$: $\{Z,A,B,C\},\{Z,A,B,D\},\{Z,A,C,D\},\{Z,B,C,D\}$

d. $Sets\ of\ 5\ nodes$: $\{Z,A,B,C,D\}$

· 왜 3개 이상의 집합도 찾아야 하는 건지?
→ Causal effect를 계산할 때 경우에 따라 다른 변수를 adjustment할 필요가 있다.
→ 예를 들어 A=a일 때 causal effect를 계산하면 condition on B를 해야 하는 경우가 존재한다.

b. (a)에서 이미 구하였다.

· $\{Z,A\},\{Z,B\},\{Z,C\},\{Z,D\}$

c. 변수들의 최소 집합들

· D

a. C : $\{C\}$

b. Z : $\{Z,A\},\{Z,B\},\{Z,X\},\{Z,W\}$

· $\{W,D\}$

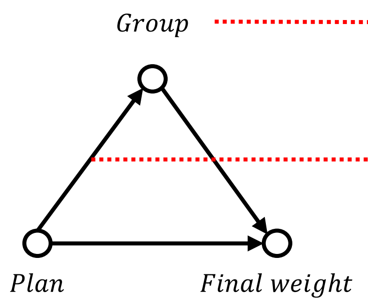
a. C, X : $\{C,X\}$

b. Z : $\{Z\}$

Study question 3.3.2 (a), (b)

a.

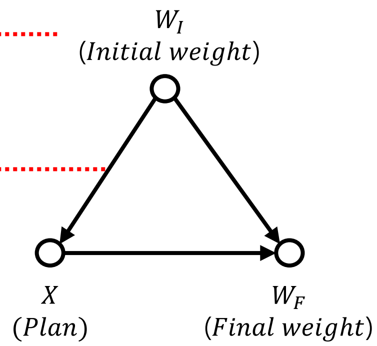
나의 풀이



· 두 번째 통계학자가 '범주화'한 것이기에 변수는 group이 아니라 처음 몸무게인 W_I가 되어야 한다.

· 식단에 따라 subgroup을 나눈다고 이해하였다.
→ 학생들의 initial weight가 식단 선택에 영향을 준다.

Solution

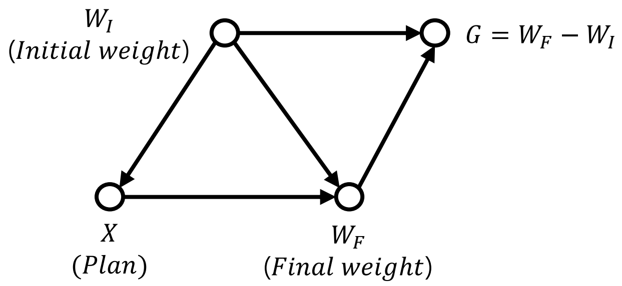


b. 두 번째 통계학자의 결과가 옳다.

→ W_I를 고려하여 *spurious path*를 막았기 때문이다.

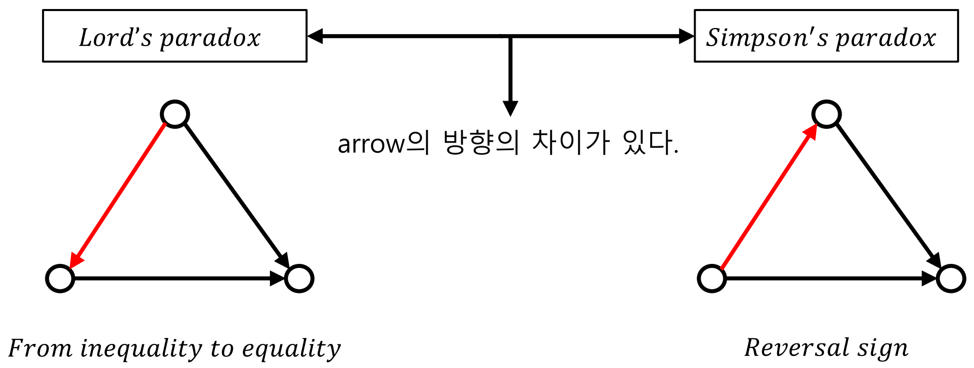
· 중요한 점은 두 몸무게의 차이 G가 주어져도 *graph model*은 변하지 않는다.

→ 따라서 G가 추가되는 것과 상관없이 W_I로 생기는 *spurious path*를 막아야 한다.

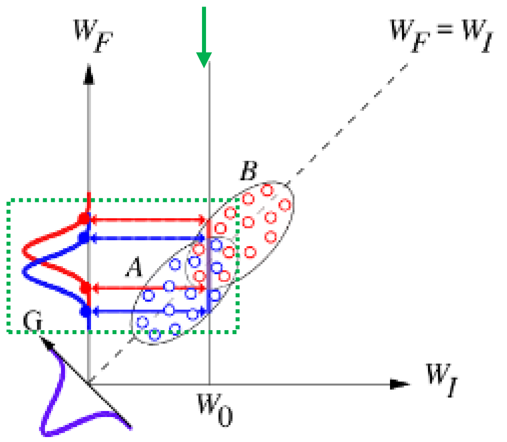


Study question 3.3.2 (c)

c. · *Lord's paradox*와 *Simpson's paradox*의 차이

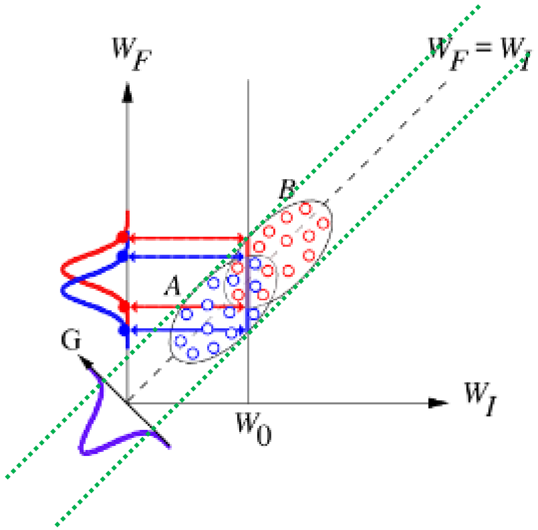


① $W_I = W_0$



W의 값이 주어진 경우 A와 B는 근소하지만 차이를 보인다.

② W가 주어지지 않는 경우

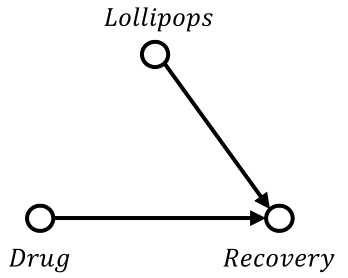


A와 B 모두 평균이 '$W_F = W_I$'인 정규분포이기에 결과가 동일해 보이게 된다.

Study question 3.3.3

a.

나의 풀이

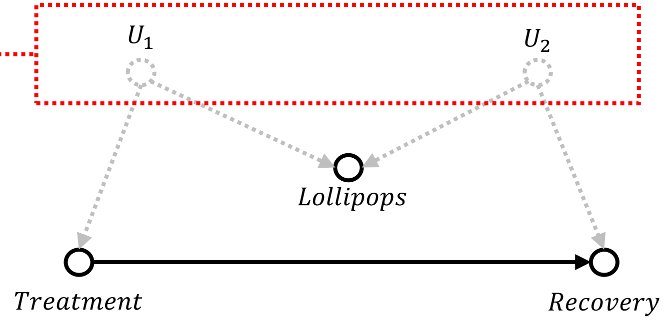


· 어떤 약을 섭취하는가와 lolipop을 받는가에 영향을 미치는 관측되지? 않은 변수가 있다?

· Recovery의 여부와 lolipop을 받는 것에 영향을 미치는 관측되지 않은 변수가 있다?

→ 두 변수 U_1와 U_2의 구조가 반드시 Fork 여야하는가?

Solution



b. Lolipop이 collider이기 때문에 별도로 조정이 필요하지 않다.

c. 조정이 필요 없기 때문에 다음과 같다.

→ $P(Y = y|do(X = x)) = P(Y = y|X = x)$

d. 변화 없다

· $X = Placebo$가 된다 하여도 $graph$에 영향을 주지 않기 때문이다.

· $Z(lollipop)$에 특정 값이 주어진다 하여도 $collider$이기 때문에 $causal\ effect$에는 차이가 없다.

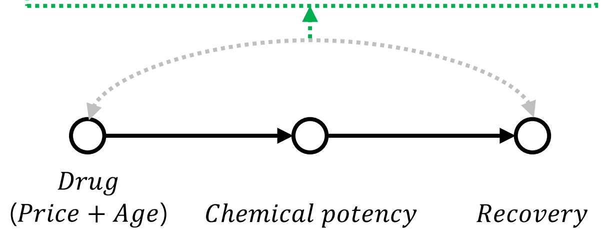
Study question 3.4.1

a. Front-door criterion을 묻는 문제이다.
→ 교재에서 설명한 부분과 완전히 동일하다.

$$P(y|do(x)) = \sum_{w} P(w|x) \sum_{x'} P(y|x', w) P(x')$$

Study question 3.4.2 (a), (b)

a. Customers는 chemical potency를 모른 채로 price를 중심으로 약을 구매한다.



*Drug*는 차례로 *chemical potency*와 *recovery*에 영향을 미친다.

· *Price*와 *Age*를 두 개의 개별 변수로 생각하여 model의 모양이 그려지지 않았다.

b. 답은 교재의 *Table* 3.2와 동일하다. 주어진 문제를 *Front – door* 문제에 적용하여 나타낼 수 있는지 확인하는 듯하다.

	***Drug*(*low price* & *old*) 400**		***Drug*(*high price* & *fresh*) 400**		***All drugs* 800**	
	Chemical low	*Chemical high*	*Chemical low*	*Chemical high*	*Chemical low*	*Chemical high*
Total patients	380	20	20	380	400	400
Recovery	323 (85%)	18 (90%)	1 (5%)	38 (10%)	324 (81%)	56 (19%)
No recovery	57 (15%)	2 (10%)	19 (95%)	342 (90%)	76 (19%)	344 (81%)

Study question 3.4.2 (c)

c. $D:\ Drug, C:\ Chemical\ potency, R:\ Recovery$

① $P(C|do(D)) = P(C|D)$

② $P(R|do(C)) = \sum_{d'} P(R|C, D = d')P(D = d')$

③ $P(R|do(D)) = \sum_{c} P(C|D) \sum_{d'} P(R|C, D = d')P(D = d')$

· $Sigma$의 대상이 $intervention\ do$를 제외한 변수라 생각하면 쉽다.

$$\cdot\ P(R|do(D = fresh)) = P(C = high|D = fresh)[P(R|C = high, D = old)P(D = old) + P(R|C = high, D = fresh)P(D = fresh)]$$
$$+P(C = low|D = fresh)[(P(R|C = low, D = old)P(D = old) + P(R|C = low, D = fresh)P(D = fresh)]$$
$$= 0.95[0.1 * 0.5 + 0.9 * 0.5] + 0.05[0.05 * 0.5 + 0.85 * 0.5]$$
$$= 0.4975$$

$$\cdot\ P(R|do(D = old)) = P(C = high|D = old)[P(R|C = high, D = old)P(D = old) + P(R|C = high, D = fresh)P(D = fresh)]$$
$$+P(C = low|D = old)[(P(R|C = low, D = old)P(D = old) + P(R|C = low, D = fresh)P(D = fresh)]$$
$$= 0.05(0.1 * 0.5 + 0.9 * 0.5) + 0.95(0.05 * 0.5 + 0.85 * 0.5)$$
$$= 0.4525$$

$$\cdot\ ACE = 0.4975 - 0.4525$$
$$= 0.045 > 0$$

Study question 3.5.1

a. $\cdot\ P(Y = y|do(X = x), C = c)$

$\rightarrow \sum_z P(Y = y|X = x, Z = z, C = c)P(Z = z)$

*Backdoor path*를 막기위해 Z의 값이 주어질 필요가 있다.

b. $\cdot\ \{A, B, C, D\}$ 중 어느 하나만 주어지면 $\{X, Z, Y\}$인 상황에서 $z - specified\ effect$를 구할 수 있다.

$\rightarrow P(Y = y|do(X = x), Z = z)$

$\rightarrow \sum_a P(Y = y|X = x, Z = z, A = a)P(A = a)$

$\cdot\ \{X, W, Y, Z\}$가 주어질 때 $z - specified$하면 W가 $front - door\ criterion$을 만족한다

$\rightarrow$ $Front - door$는 $W\ to\ Y$와 $X\ to\ W$의 $causal\ effect$를 이용하는 것인데 Z와 어떤 상관이 있는지 모르겠다??

c. $\cdot\ g(Z) \rightarrow \begin{cases} 0, Z \le 2 \\ 1, Z > 2 \end{cases}$

X의 값 x를 Z에 대한 함수로 표현한다.

$\cdot\ P(Y = y|do(X = g(Z))) = \sum_z P\big(Y = y\big|do\big(X = g(z)\big), Z = z\big)P(Z = z)$

$= P(Y = y|do(X = 0), Z = 1)P(Z = 1)$

$+P(Y = y|do(X = 0), Z = 2)P(Z = 2)$

$+P(Y = y|do(X = 1), Z = 3)P(Z = 3)$

$+P(Y = y|do(X = 1), Z = 4)P(Z = 4)$

$+P(Y = y|do(X = 1), Z = 5)P(Z = 5)$

이 부분을 (b)에 적용해서 풀이하는 듯하다.

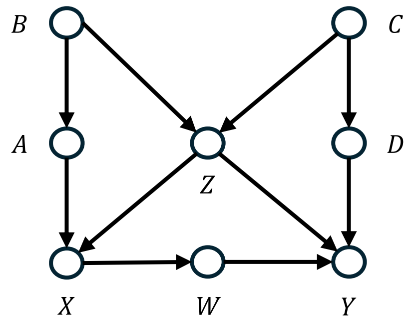
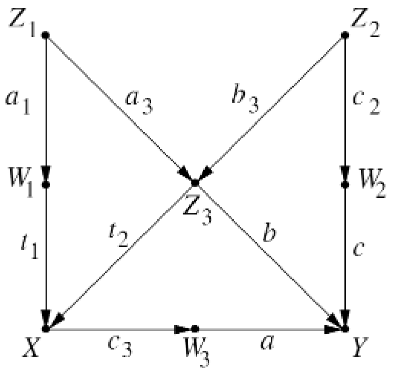


Figure 3.8

Study question 3.8.1 (a), (b), (c)



a. 모델을 검증하기 위해서는 조건부 독립의 여부를 확인해야 한다.

· $W_3 = r_X X + r_{W_1} W_1$ (r_{W_1}=0, chain)

· $W_1 = r_{Z_1} Z_1 + r_{Z_3} Z_3$ (r_{Z_3}=0, fork)

· $Y = r_{Z_1} Z_1 + r_{W_1} W_1 + r_{Z_2} Z_2 + r_{Z_3} Z_3$ (r_{Z_1}=0, chain)

b. · $W_3 = r_X X + r_{Z_3} Z_3$ (r_{Z_3}=0, chain)

c. ① Parent nodes만을 이용해 regression을 하면 model parameter를 이용한 회귀가 가능하다.

· $a = r_{W_3},\ b = r_{Z_3},\ c = rW_2$

$$Y = r_{W_3} W_3 + r_{Z_3} Z_3 + r_{W_2} W_2$$

· $a_1 = r_{Z_1}$

$$W_1 = r_{Z_1} Z_1$$

· $a_3 = r_{Z_1},\ b_3 = r_{Z_2}$

$$Z_3 = r_{Z_1} Z_1 + r_{Z_2} Z_2$$

· $c_2 = r_{Z_2}$

$$W_2 = r_{Z_2} Z_2$$

· $c_3 = r_X$

$$W_3 = r_X X$$

· $t_1 = r_{W_1},\ t_2 = r_{Z_3}$

$$X = r_{W_1} W_1 + r_{Z_3} Z_3$$

② 3.8.3 → The Regression Rule for Identification.

· $a = r_{W_2},\ b = r_{Z_3},\ c = r_{W_2}$

$$Y = r_{W_3} W_3 + r_{Z_3} Z_3 + r_{W_2} W_2$$

$Or,\ a = r_{W_2},\ b = r_{Z_3},\ c = r_{Z_2}$

$$Y = r_{W_3} W_3 + r_{Z_3} Z_3 + r_{Z_2} Z_2$$

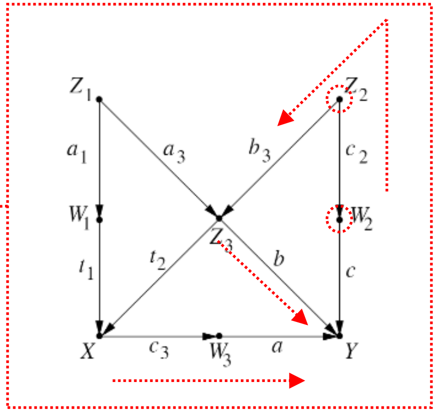
· $t_1 = r_{W_1}$

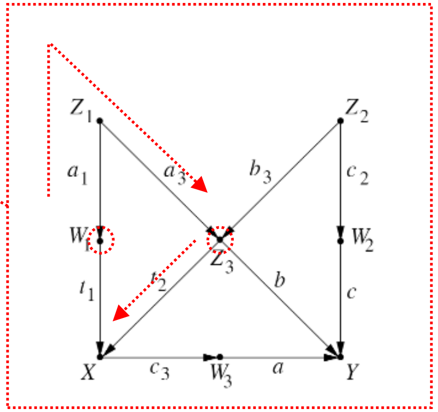
$$X = r_{W_1} W_1 + r_{Z_2} Z_2$$

$Or\ t_2 = r_{W_1}$

$$X = r_{W_1} W_1 + r_{Z_2} Z_3$$

- 직접적인 경로(edge)를 차단하고, d-separated하게 만드는 변수들이 있다면 이를 통해 α를 계산할 수 있음을 의미하는 듯하다(?).





Study question 3.8.1 (d), (e), (f), (g)

d. ① c_3 & a는 주어진 정보에서 알 수 있다.

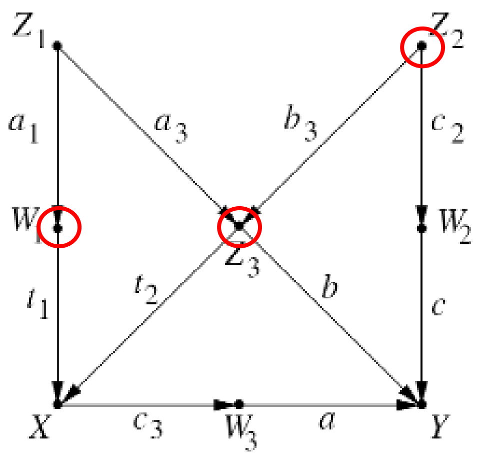
· $W_3 = r_X X + U \rightarrow c_3$

· $a = r_{YW_3 \cdot X}$ ($W_3 \rightarrow Y$는 X로 backdoor path를 막을 수 있다.

② W_3로 인해 Font-door criterion을 만족한다.

$\rightarrow a * c_3$

e. $\{W_X, W_2, W_3, Y\}$, 붉은 색 변수들로 인해 다른 변수들과 independent 하다.



f. 문제의 핵심은 새로운 regressor의 추가가 종속변수와 independent 해야 한다.

· $W_1 = r_{Z_3} Z_3 + r_X X + U$

$W_1 = r_{Z_3} Z_3 + r_X X + r_{W_3} W_3 + U \rightarrow$ Dependent한 path가 없기 때문에 invariant 하다.

$W_1 = r_{Z_3} Z_3 + r_X X + r_Y Y + r_{Z_2} Z_2 + +U \rightarrow$ Dependent한 path가 생기기 때문에 coefficient가 변한다.

g. '3.8.3 → Instrumental variable'을 확인하면 풀이 과정을 알 수 있다.

① Z_1을 instrumental variable로 하여 Y를 Z_1의 회귀로 표현한다.

· Backdoor path를 막기위해 W_1을 condition 해야 한다.

· Unmeasured의 문제는 Y를 Z_3로 회귀하면 '$Z_3 \leftarrow Z_2 \rightarrow W_2 \rightarrow Y$'로 b를 알 수 없다.

→ 반면 Z_1은 Z_3가 주어지지 않으면 Z_2와 independent하다.

→ '$Z_1 \rightarrow Z_3 \rightarrow Y$' 가능

· 따라서 회귀식은 '$Y = r_{Z_1} Z_1 + r_{W_1} W_1 + U$'이 된다.

→ Z_1의 회귀 경로는 두 가지가 존재 한다. 하나는 '$Z_1 \rightarrow Z_3 \rightarrow Y$'이고, 다른 하나는 '$Z_1 \rightarrow Z_3 \rightarrow X \rightarrow W_3 \rightarrow Y$' 이다.

→ 따라서 $r_{Z_1} = a_3 b + a_3 t_2 c_3 a = a_3(b + t_2 c_3 a)$ 이 된다.

② Z_3을 Z_1으로 회귀

· $Z_3 = r_{Z_1'} Z_1' + U$

$\rightarrow r_{Z_1'} = a_3$

③ Y에 대한 Z_3의 회귀계수 계산

· $\frac{r_{Z_1}}{r_{Z_1'}} = \frac{a_3(b + t_2 c_3 a)}{a_3} = b + t_2 c_3 a$

→ t_2, c_3, a는 모두 구할 수 있기에 식을 정리하면 b를 구할 수 있다.