Study questions

1.2.1

(a) Correlation does not imply causation. coincidence

(b) Reverse causation

(c) Reverse causation

1.2.2

|  |  |  |  |
| --- | --- | --- | --- |
|  | R | L | Avg |
| Frank | 99/100 | 60/100 | 159/200 |
| Tim | 1/1 | 2/3 | 3/4 |

1.2.3

(a)

The same as (b), just change the difficulty to the size of stone.

(b)

The difficulty of surgery is a confounder for the choice of doctor and the rate of success of the surgery (plot the classical confounder graph). The difficulty of surgery affects the choice of doctor and also chances of recovery as more difficult cases can have less chance of success. Therefore, to make a causal conclusion, we need to consult the segregated data, by conditioning on the level of difficulty.

1.2.4

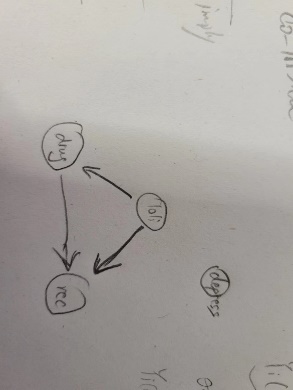
(a)

Yes, the drug is beneficial to the population as a whole.

(b)

No.

(c)



(d)

(e) Yes. Because if loli was given after the treatment, we cannot assume the effect that loli to the treatment but recovery only.

1.3.1

Lollipop: yes/no

Treatment: yes/no

Recovery: yes/no

1.3.2

(a) (231+ 189) / (112+231+595+242+136+189+763+172) = 0.1721311475409836

(b) (231+ 136+189+763+172) / (112+231+595+242+136+189+763+172) = 0.6110655737704918

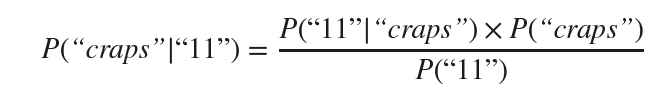
(c) (189)/(136+189+763+172) = 0.15

(d) 189/( 231+ 189) = 0.45

1.3.3

(a)





0.5135135135135135

(b)

No need.

1.3.4

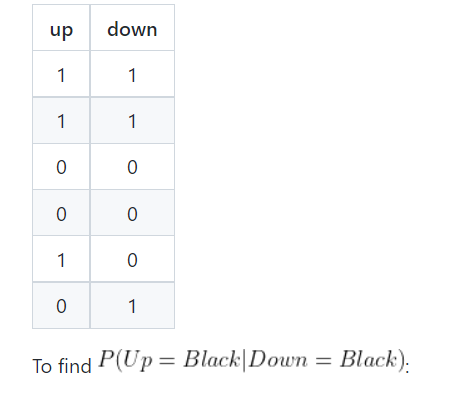
(a)

Intuitively: because if one face is black, you rule out the possibility of being card2. Only two options are left.

However, maybe you should consider card1 twice because it could have landed in both sides.

(b)

(c)



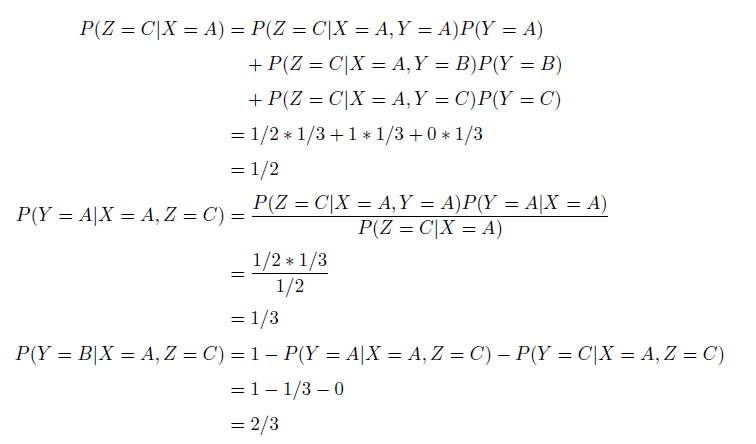
1.3.5

Let A, B, C be the three doors.

Let 𝑋 indicate the door chosen by the player, 𝑌 indicate the door hiding the car, and 𝑍

indicate the door opened by the host.

𝑃(𝑌 = 𝐴|𝑋 = 𝐴,𝑍 = 𝐶) < 𝑃(𝑌 = 𝐵|𝑋 = 𝐴,𝑍 = 𝐶)



1.3.6

PASS

1.3.7

(a)

P(x)=0.75

P(y)=0.5

P(x,y) = 0.25

P(y|x) = 1/3

P(x|y) = ½

(b)

E[X] =0.75

E[Y] = 0.5

E[Y|X=x] = 1/3

E[X|Y=y] = ½

Var(X) = (1-3/4)^2\*3/4 + (0-3/4)^2\*1/4 =3/16

Var(Y) = (1-1/2)^2 \* ½ + (0-1/2)^2 \* ½ = ¼

Cov(X,Y) = 1\*1/4 + 0\*3/4 – ¾\*1/2 = -1/8

P\_xy = - 1/sqrt(3)

(c)

½

(d)

1/3

(3)

No

y

1.3.8

(a)

E[X] = 7/2

E[Y] = 7

E[Y|X=x] = x + E[x]

E[X|Y=y] = y/2

Var(X) = (1-7/2)^2\* 1/6 + … + (6-7/2)^2\* 1/6 = 2.9166

Var(Y) = 5.833333

Cov(x,y) =

1.3.9

(a)

Y=a+bx

E[Y] = E[a] + E[bx]

E[XY] = aE[y] + bE[XY]

(b)

Cov(z,x) =

Cov(z,y) =

Q求解2元方程

1.4.1

(a)W, Y

(b)X, W, Y

(c) Y, Z

(d)Y, Z, T

(e) x-y-t

x-y-z-t

x-y-w-z-t

x-w-y-t

x-w-z-t

x-w-y-z-t

x-w-z-y-t

(f)

x-y-t

x-y-z-t

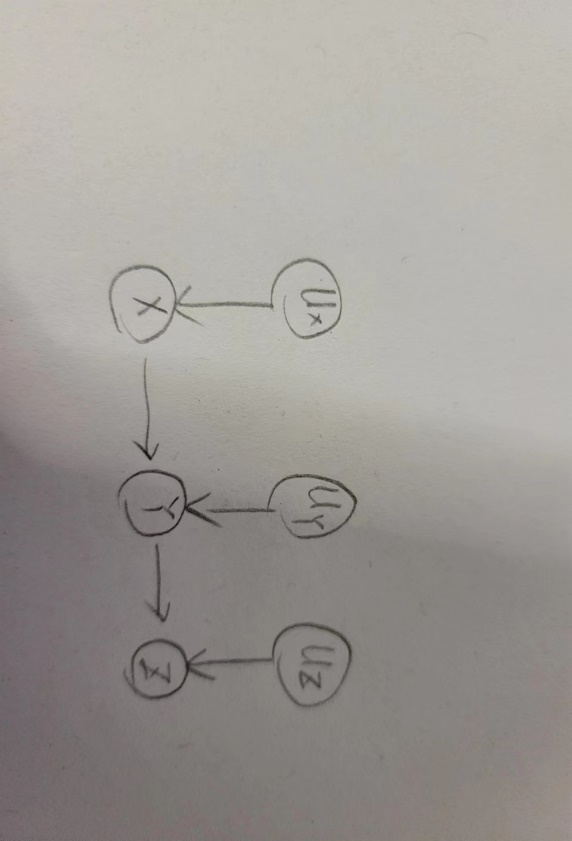
x-w-y-t

x-w-z-t

x-w-y-z-t

1.5.1

(a)



(b)

3/16

(c)

1/16

(d)

3/16

(e)

i.6

ii. use linear regression

1.5.2

(a)

P(x,y,z) = p(y|x,z)P(z)P(x|z)

P(x,y) = sum\_z(P(x,y,z))

1.5.3

PASS

1.5.4

(a)

The Monty Hall problem has two exogenous variables (participant’s choice and the position of the car) and one endogenous variable (Monty’s choice). Calling the exogenous variables U1 and U2, respectively, and the endogenous variable V1:

U = {U1, U2}; V = {V1}; f = {f1};

V1 = f1(U1, U2)

(b)

P(U1,U2,V1) = P(U1) P(U2)P(V1|U1,U2)

2.3.1

(a)

('u', 'x', {'v', 'r'})

('u', 's', {'v', 'r'})

('u', 'y', {'v', 'r'})

('x', 's', {'v', 'r'})

('x', 't', {'v', 'r'})

('x', 'y', {'v', 'r'})

('s', 'y', {'v', 'r'})

('t', 'y', {'v', 'r'})

(b)

path.get\_all\_independence\_relationships()

(c)

('u', 'x', {'r', 'p'})

('x', 's', {'r', 'p'})

('x', 't', {'r', 'p'})

('x', 'y', {'r', 'p'})

('x', 'v', {'r', 'p'})

(d)

path.get\_all\_independence\_relationships()

(e)

u,s,r,v

(f)

B, c, f

2.4.1

(a)

[('w', 'z3', {'x'}),

('w', 'z3', {'x', 'z1'}),

('w', 'z3', {'x', 'z2'}),

('w', 'z3', {'x', 'z1', 'z2'}),

('w', 'z2', {'x'}),

('w', 'z2', {'z1', 'z3'}),

('w', 'z2', {'x', 'z3'}),

('w', 'z2', {'x', 'z1'}),

('w', 'z2', {'x', 'z1', 'z3'}),

('w', 'z1', {'x'}),

('w', 'z1', {'x', 'z3'}),

('w', 'z1', {'x', 'z2'}),

('w', 'z1', {'x', 'z2', 'z3'}),

('w', 'z1', {'x', 'y', 'z2', 'z3'}),

('x', 'z2', {'z1', 'z3'}),

('x', 'z2', {'w', 'z1', 'z3'}),

('x', 'z2', {'w', 'y', 'z1', 'z3'}),

('x', 'y', {'w', 'z1', 'z3'}),

('x', 'y', {'w', 'z2', 'z3'}),

('x', 'y', {'w', 'z1', 'z2', 'z3'}),

('z2', 'z1', set()),

('y', 'z1', {'x', 'z2', 'z3'}),

('y', 'z1', {'w', 'z2', 'z3'}),

('y', 'z1', {'w', 'x', 'z2', 'z3'})]

(b)

('w', 'z3', {'x'})

('w', 'z3', {'z1', 'x'})

('w', 'z1', {'x'})

('w', 'z1', {'z3', 'x'})

(c)

('w', 'z1', {'z3', 'y', 'x', 'z2'})

('x', 'z2', {'z3', 'y', 'z1', 'w'})

('x', 'y', {'z3', 'z1', 'w', 'z2'})

('y', 'z1', {'z3', 'x', 'w', 'z2'})

(d)

W:

(e)

W, z2, z3

(f)

W,y,z1,z3

(g)

Yes, because z2 and w are dependent when conditioning on z3, so no slope will be zero. As a matter of fact, regression with high-order (more valid variables) tends to have better performance than low-order.

2.5.1

(a)

All edges are connected to a collider or would become a collider if they were inverted. Therefore, this model is unique.

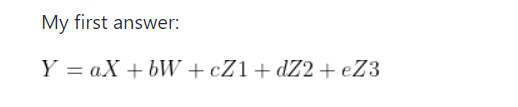
(b)

Since the graph in 2.9. is unique, there is no equivalent one.

(c)

Detect which arrow can be reversed.

(d)



Since X and Z1 are separated from Y, we can infer that if the coefficients associated with these variables (a and c) are nonzero, the model is wrong. **This answer provides a valid answer, but it is not how one should solve it.**

**(e)**



Because [('w', 'z3', {'x'}),

(f)

Because

[('w', 'z3', {'x'}),

('w', 'z3', {'x', 'z1'}),

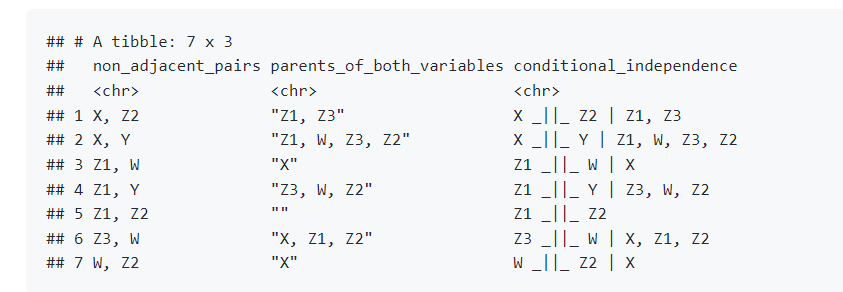
('w', 'z3', {'x', 'z2'}),

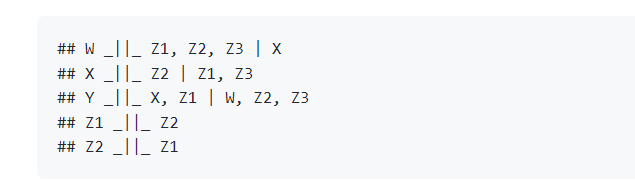
('w', 'z3', {'x', 'z1', 'z2'}),

If x is not measured, then all other regression without x doesn’t have a coefficient that is zero, which we could leverage to test whether the model is wrong. I.e. omitting x, we cannot find independent relation between z3 and any other variables.

(g)

According to Kline (2015, p. 173): The basis set for a DAG includes the smallest number of conditional independences that imply all others (if any) located by the d-separation criterion. **The size of the basis set equals the number of pairs of nonadjacent variables that can be d-separated.** Any conditional independences beyond this number are implied by the basis set; that is, they are redundant, so there is no need to test them all.





**find the basis set**: List each unique pair of nonadjacent variables in the graph that can be d-separated. Next, condition on the parents of both variables in each pair. The corresponding set of conditional independences is a basis set.

3.2.1

(a)

3.3.1

(a)

frozenset({frozenset({'a', 'b', 'z'}),

frozenset({'c', 'd', 'z'}),

frozenset({'a', 'z'}),

frozenset({'a', 'b', 'c', 'z'}),

frozenset({'b', 'd', 'z'}),

frozenset({'c', 'z'}),

frozenset({'a', 'c', 'z'}),

frozenset({'d', 'z'}),

frozenset({'b', 'z'}),

frozenset({'b', 'c', 'z'}),

frozenset({'b', 'c', 'd', 'z'}),

frozenset({'a', 'd', 'z'}),

frozenset({'a', 'c', 'd', 'z'}),

frozenset({'a', 'b', 'd', 'z'}),

frozenset({'a', 'b', 'c', 'd', 'z'})})

(b)

(a,z)

(b,z)

(c,z)

(d,z)

(c)

[frozenset({'c', 'x'}),

frozenset({'c', 'z'}),

frozenset({'x', 'z'}),

frozenset({'a', 'z'}),

frozenset({'b', 'z'})]

3.3.2

(a)

WF<-WI->Plan->WF

(b)

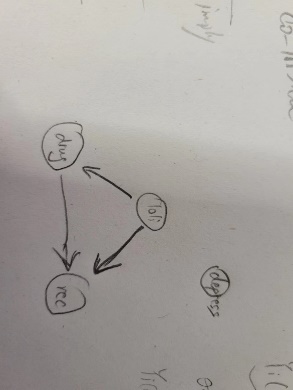
Second

(c)

Left hand & right hand example.

3.3.3

(a)



(b)

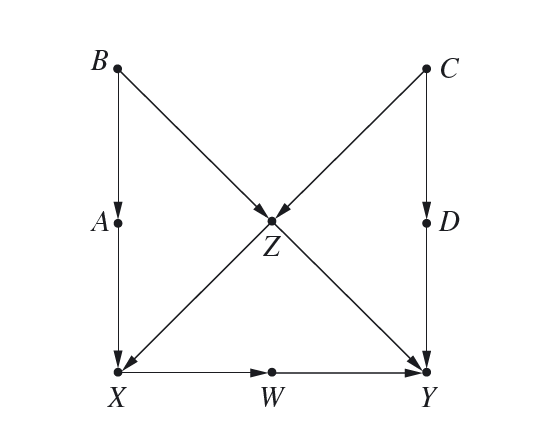
Lollipop

(c)

P(recovery| do(drug)) = sum\_lollipop (p(recovery| drug, lollipop)P(lollipop))

(d)

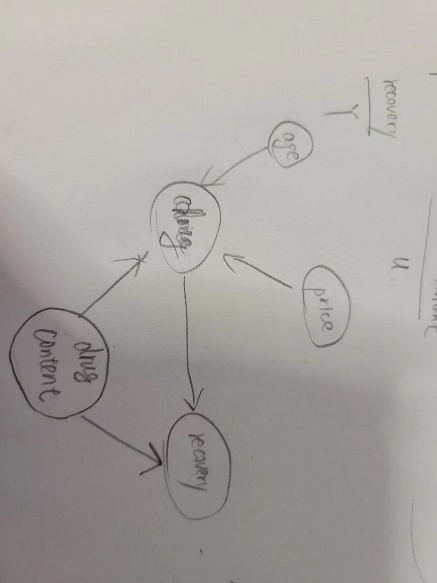
3.4.1



W. Front door

3.4.2

(a)



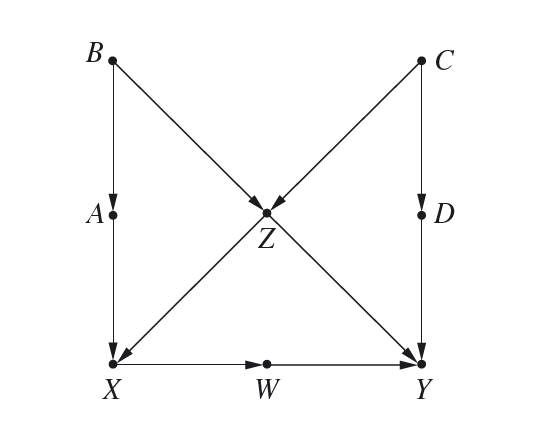
(b)

|  |  |  |  |
| --- | --- | --- | --- |
|  | Fresh drug (expensive) | Aged drug (cheap) |  |
| Have active ingredient | 10/38 | 1/4 | 11/42 |
| Not have active ingredient | 2/2 | 22/76 | 13/78 |
| Combine data | 12/40 recovery | 23/80 recovery |  |

(c)

P(Y=y|do(T=t)) = P(Y=y|)

3.5.1



(a)

P(Y=y| do(X=x), C=c) = sum(P(Y=y| X=x, C=c, Z=z) P(Z=z| C=c) )

(b)

{X,Y,Z,A}, {X,Y,Z,B}, {X,Y,Z,C}, {X,Y,Z,D}

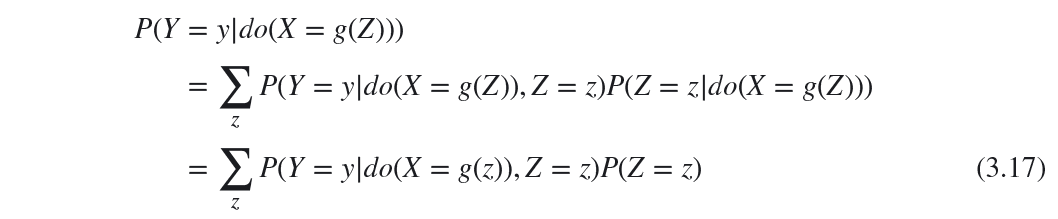
For example. We use {X,Y,Z,A}

P(Y=y | do(X=x), Z=z) = sum\_a( P(Y=y|do(X=x), Z=z, A=a) P(A=a|Z=z) )

(c)

x= 0 if z <=2

x=1 if z > 2



3.8.1

(a)

(b)

We could use back door criteria to estimate c3 by using

We can use front door criteria to estimate a\*c3

(c)

Y =

(d)

A, c3

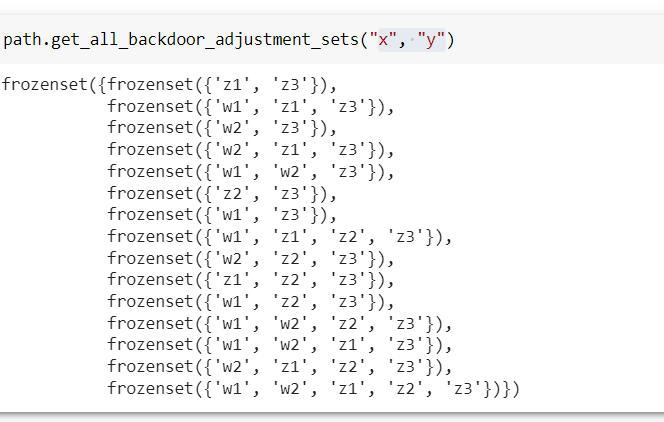
Yes. Because TE = E[Y|do(X=1)] – E[Y|do(X=0)]

E[Y|do(X=0)] use front door.

(e)

W2, Z3

(f)



求backdoor adjustment set 包含关系

(g)

How to turn z1 into an instrument variable?

4.3.1

(a)

(b)

4.3.2

(a)

Regression (X, H) get a

4.4.1

(a)

E[Yx|x′] = (E[Yx] - E[Yx|x]P(x))/ P(x′)

(b)

(c)

4.4.2

(a)

E[Y\_x|X=x’]

Y=cancer

X= smoke

X’ = not smoke

(b)

The peer data who smoked and the probability to get cancer, and who doesn’t smoke and the probability to get cancer.

(c)

4.5.1

4.5.2

(a)

TE = b1r1+b2

NIE = b1

NDE=b2

(b)

4.5.3

(a)

(b)

4.5.4