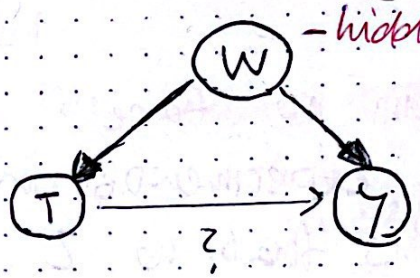
Lecture 1

Methods for Causal Inference

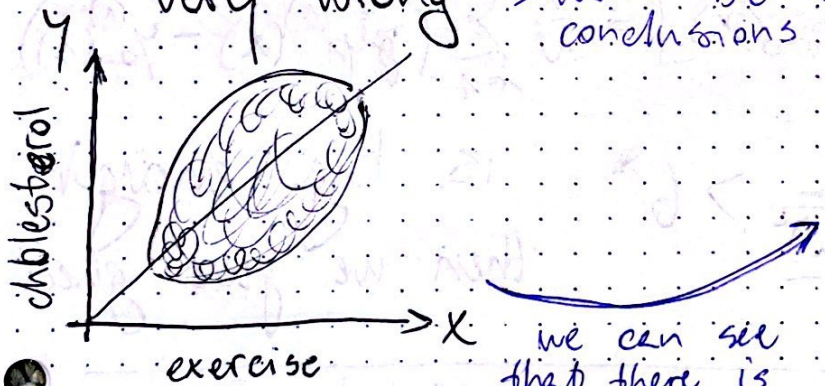
- **Reverse causation**: wind blows ⟷ wind turbines rotate faster

if we don't see the bigger picture - knowing what causes what - then we may conclude the reverse causation!

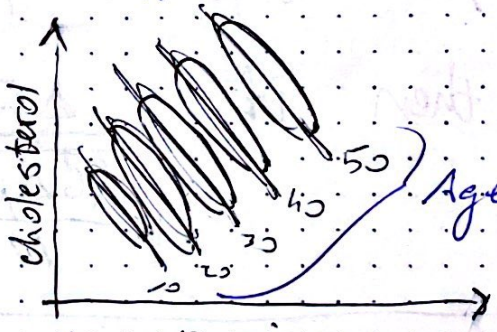
Confounding factor - hidden variable



- **Simpson's Paradox** - concluding causality from purely associational measure (correlation) can be very wrong → would be better not to draw any conclusions at all!!

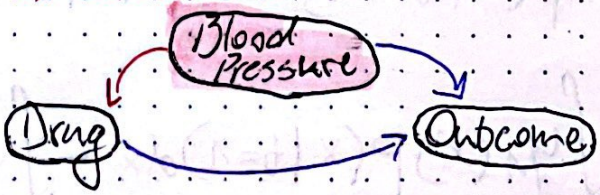


we can see that there is a factor we haven't accounted for



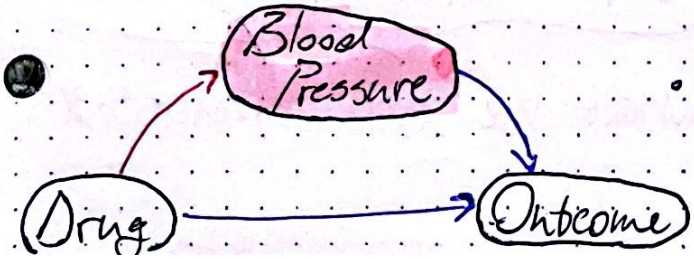
Consider all variables affecting the system of interest and the role they play

*scenario 1



- blood pressure is a confounder here

*scenario 2



- blood pressure is a mediator

T - treatment
Y - outcome
X - confounders
U - unobservable confounders

$$\mathbb{E}[Y_{b=1}(x) - Y_{b=0}(x)] =$$

(expected value of outcome Y when we apply the treatment vs. when we don't given some confounders (= features)

$$= \int (Y_1(x) - Y_0(x)) p(x) dx$$

(probability distribution of subjects given confounder X (like age or sex or sth)

The problem we face when doing experiments on humans is that no 2 humans are genetically the same ⇒ let's take average across the whole population

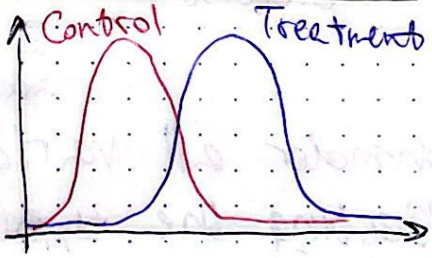
$$\Delta\bar{\mu} = \bar{\mathbb{E}}[Y_{b=1}(x) - Y_{b=0}(x)] = \frac{1}{N}\sum_{i=1}^{N}\left(Y_1^{(i)}(x) - Y_0^{(i)}(x)\right)$$

then if $\dfrac{\Delta\bar{\mu}}{\sqrt{\frac{(\sigma\Delta\mu)^2}{N}}} > b^*$ is big enough then we gucci gucci

---

~~Oops~~ What's wrong?

$$p(x|b=1) \neq p(x|b=0) \Rightarrow$$



$$\int Y_1(x) p(x|b=1) dx - \int Y_0(x) p(x|b=0)$$

$$\neq \int (Y_1(x) - Y_0(x)) p(x) dx$$

if we naively measure it, then we will get false result!

what we are interested in

possible solution is STRATIFICATION
↳ measure outcome within each of the groups seperately

$$\mathbb{E}(\text{healed} \mid t=1) = \mathbb{E}(\text{healed} \mid t=1, \text{young})\, p(\text{young}) + \mathbb{E}(\text{healed} \mid t=1, \text{old})\, p(\text{old})$$

Disadvantages:

1. all possible outcomes need to be observed
2. samples can become not representative

- 2 main frameworks:

↳ causal estimation!

#1 Potential Outcomes => Rubin

#2 Structural Causal Models => Pearl

↳ causal discovery

Lecture 2

Probability axioms for two mutually exclusive events:

$P(A \text{ or } B) = P(A) + P(B)$

$P(A \text{ and } B) = P(A) \cdot P(B)$

$P(A) = P(A, \text{not } B) + P(A, B)$

Conditional Probability

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

Bayes rule

$$P(X_i|Y) = \frac{P(Y|X_i)P(X_i)}{P(Y)}$$

$$= \frac{P(X_i)P(Y|X_i)}{P(X_1)P(Y|X_1) + \ldots + P(X_n)P(Y|X_n)}$$

Incorporating knowledge about the process that generated the data ⇒ **FIRST STEP TOWARDS CAUSAL INFERENCE**

Two events are said to be mutually exclusive when their occurrence is not simultaneous

Two events are said to be independent when occurrence of one cannot control occurrence of other

$$P(X, Y) = P(X)P(Y)$$

$$P(X|Y) = P(X) \quad \forall P(Y) \neq 0!$$

Expected Value – sum over the probability distribution

$$\mathbb{E}[X] = \sum_x x P(X=x)$$

$\mathbb{E}[X]$: random variable; x: values RV can take

for continuous variables $\Sigma \rightarrow \int$:

$$\mathbb{E}[X] = \int x P(x)\, dx$$

Variance

$$\text{var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \sum_x (X - \mathbb{E}[X])^2 P_X(x)$$

↑ how spread out are the values!

Covariance

$$\sigma_{XY} = \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right]$$

↑ degree to which X and Y covary

when normalised, we get a coefficient:

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$ ⇐ PEARSON CORRELATION

$\rho_{XY} \in [-1, 1]$ but $\rho_{XY} = 0$, when X and Y are independent

↳ but ρ_{XY} does not imply independence.

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

- R^2 coefficient ~ a measure for goodness of a fit

$$R^2 = 1 - \frac{\sum_i (y_i - f_i)}{\sum_i (y_i - \bar{y})}$$

(f_i: prediction; y_i: actual point; $\bar{y}$: mean)

- if fit is perfect, then $R^2 = 1$ and $R^2 = 0$ implies fit is no better than baseline $\bar{y}$

Lecture 3

In linear regression modeling, optimal solution is reached when $\beta = \frac{Cov[X,Y]}{Var[X]}$ in:

$$y = \alpha + \beta X$$

- $\beta > 0$ – positive correlation
- $\beta < 0$ – negative correlation
- $\beta = 0$ – no linear correlation

Multi regression

y on vars X_1 & X_2:

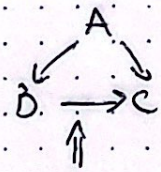
$$y = \alpha + \beta_1 X_1 + \beta_2 X_2$$

V-fold

① split the data into V blocks:  and in each of V iterations, one block will be used for testing while rest for training

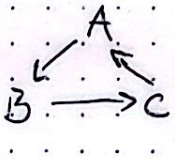
Directed Graphs

Acyclic



DAGs

Cyclic



STRUCTURAL CAUSAL MODELS

- 2 vars U & V and set of functions of

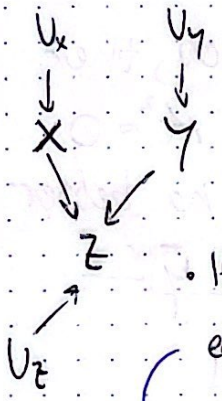
U - exogenous variable:
↳ external to the model

V - endogenous:
↳ descendant of at least one exogenous variable

PRODUCT DECOMPOSITION RULE

$X \to Y \to Z$ $\quad P(X=x, Y=y, Z=z) =$

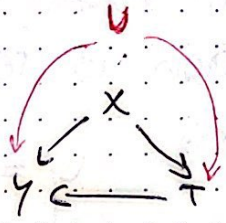
$= P(X=x) P(Y=y \mid X=x) P(Z=z \mid Y=y)$



$P(X,Y,Z) = P(Z \mid X,Y) P(X) P(Y)$

- If $U_X \perp\!\!\!\perp U_Y \perp\!\!\!\perp U_Z$ – their outer noises are independent of each other, then we can drop them from our model

↳ are not confounders



* fucked!

Lecture 4

Potential Outcomes Framework [Rubin-Neyman]

- $y^{(i)}_{observed} = t^{(i)} y^{(i)}_1 + (1 - t^{(i)}) y^{(i)}_0 = \begin{cases} y^{(i)}_0 & \text{if } t^{(i)} = 0 \\ y^{(i)}_1 & \text{if } t^{(i)} = 1 \end{cases}$

but in the real world one patient cannot undergo both scenarios → that's why we introduce **counterfactuals**

$$y^{(i)}_{CF} = (1 - t^{(i)}) y^{(i)}_1 + t^{(i)} y^{(i)}_0$$ # missing outcome

now ideally we want to introduce a measure

- $\tau^{(i)} = y^{(i)}_1 - y^{(i)}_0$ telling us what is the **treatment effect** for that individual person.

$$\tau = \hat{E}[\tau^{(i)}] = \hat{E}[y^{(i)}_1 - y^{(i)}_0] = \frac{1}{N} \sum_{i=0}^{N} (y^{(i)}_1 - y^{(i)}_0)$$

Regression Adjustment

fitting $y = \beta_x X + \beta_\tau T + \varepsilon$

- $T = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ \dots & \dots \\ 0 & 1 \\ 0 & 1 \end{pmatrix}$ $X = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ \dots & \dots \\ 1 & 0 \\ 0 & 1 \end{pmatrix}$ → $\begin{pmatrix} y^{(1)} \\ y^{(2)} \\ \dots \\ y^{(N-1)} \\ y^{(N)} \end{pmatrix} = \begin{pmatrix} \beta_{b=0} + \beta_{x=y} \\ \beta_{b=0} + \beta_{x=0} \\ \dots \\ \beta_{b=1} + \beta_{x=y} \\ \beta_{b=1} + \beta_{x=0} \end{pmatrix}$

Control Drug Young Old

Assumptions: **overlap** and **exhaustivity**

- $\tau = \hat{E}[\tau^{(i)}] = \hat{E}[y^{(i)}_1 - y^{(i)}_0] = \frac{1}{N} \sum_{i=0}^{N} (y^{(i)}_1 - y^{(i)}_0)$

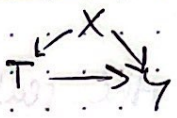
since we computed the β we can easily compute the counterfactuals

Another way of computing the counterfactuals is with **MATCHING**, where we try to create clone/twin for each individual ⇒ try to find most similar unit from sample

---

Lecture 5

Adjustment Formula



$$\mathbb{E}[Y_1 - Y_0 | X] = \mathbb{E}[Y_1|X] - \mathbb{E}[Y_0|X] =$$

CATE

$$= \mathbb{E}[Y_1 | T=1, X] - \mathbb{E}[Y_0 | T=0, X] =$$

$$= \mathbb{E}[Y_1 | T=1, X] - \mathbb{E}[Y | T=0, X]$$

ATE

$$\mathbb{E}[Y_1 - Y_0] = \mathbb{E}_X\left[\mathbb{E}[Y_1 - Y_0 | X]\right] = \mathbb{E}_X\left[\mathbb{E}[Y|T=1,X] - \mathbb{E}[Y|T=0,X]\right]$$

adjustment formula

Fit model for $Q(T,X) = \mathbb{E}[Y|T,X]$

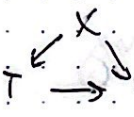
- under **linear assumptions**:

$$\mathbb{E}[Y|T,X] = \alpha_0 + \beta_X X + \beta_T T + \varepsilon; \quad \mathbb{E}[\varepsilon] = 0$$

$$ATE = \mathbb{E}_X\left[\mathbb{E}[Y|T=1,X] - \mathbb{E}[Y|T=0,X]\right] =$$

$$= (\alpha_0 + \beta_X X + \beta_T \cdot \overset{(T=1)}{1}) - (\alpha_0 + \beta_X X + \beta_T \cdot 0) =$$

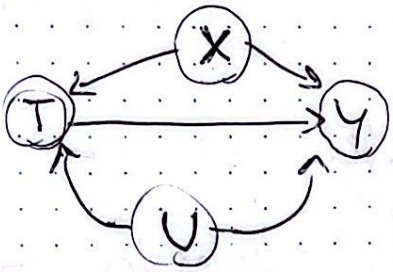
$$= \beta_T$$

which is correct! for 

if we condition on X, then Y only depends on T!

Instrumental variable

U - unobserved confounder → conditioning on X alone would not result in a randomised treatment



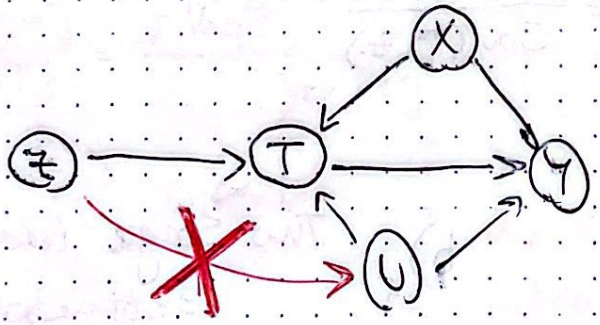
↳ we don't have access to U

What to do? Introduce new INSTRUMENTAL variable!

Example:
- T - smoking during pregnancy
- Y - birthweight of child
- X - parity / mom's age / weight ...
- U - smth (no access)

Randomise Z (intention-to-treat): either receive encouragement to stop smoking ($Z=1$) or receive usual care ($Z=0$)

$$\mathbb{E}(Y|Z=1) - \mathbb{E}(Y|Z=0)$$ ← effect Z on Y



wants to use this to get effect T on Y

any effect of Z on Y is only through T

$$(Y|Z=1,T) = (Y|Z=0,T)$$

0

We want ATE (average treatment effect):

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

but we can't get this from our graph directly cuz of U!

However, we can calculate:

$$\tau = \frac{\mathbb{E}[(Y|z=1)-(Y|z=0)]}{\mathbb{E}[(T|z=1)-(T|z=0)]}$$

which is exactly ETA!

* z was decided randomly to subjects

#1 The Wald Estimator (for binary variables)

$$\hat{\tau} = \frac{\frac{1}{n_{z=1}}\sum_{i\in z=1} y^{(i)} - \frac{1}{n_{z=0}}\sum_{i\in z=0} y^{(i)}}{\frac{1}{n_{z=1}}\sum_{i\in z=1} T^{(i)} - \frac{1}{n_{z=1}}\sum_{i\in z=0} T^{(i)}}$$

#2 IV Estimator

Linear case: $\tau = \frac{Cov(Y,z)}{Cov(T,z)}$

a) $\hat{\tau} = \frac{\hat{Cov}(Y,z)}{\hat{Cov}(T,z)}$

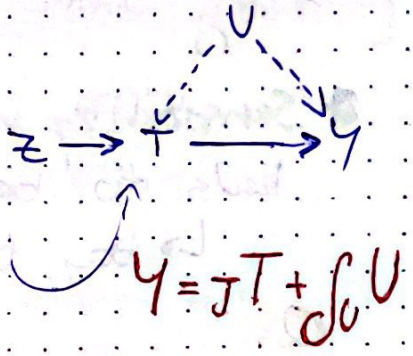
b) Two-Stage least-Square Estimator

$$Cov(Y,Z) = \mathbb{E}[YZ] - \mathbb{E}[Y]\mathbb{E}[Z] = \ldots =$$

$$= J\,Cov(T,Z) + \beta_U\,Cov(U,Z) =$$

$$= J\,Cov(T,Z)$$

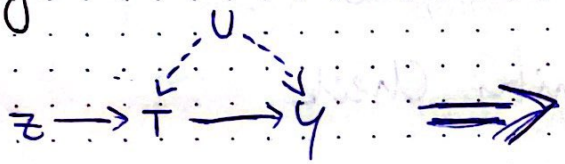
* but since there is no covariation between U and Z this is 0!

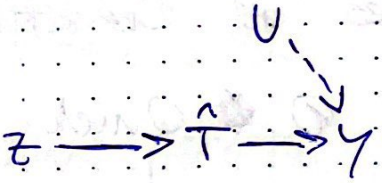


$$Y = JT + \beta_U U$$

b) [1] Estimate $\mathbb{E}[T|Z]$ to obtain $\hat{T}$ in subspace Z

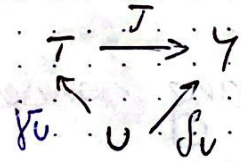
[2] Estimate $\mathbb{E}[Y|\hat{T}]$, to obtain $\hat{J}$, which is fitted coefficient in front of $\hat{T}$ in this regression





$$Y = \hat{J}\hat{T} + \beta_U U$$

Lecture 7. In the case stated above ression leads to bias!



naive regr-

$$Y = JT + \beta_U U$$

$$T = \gamma_U U$$

If we perform naive regression:

do these ?

$$\frac{Cov[T,Y]}{Var[T]} = \frac{J\,Var[T] + \beta_U\,Cov(T,U)}{Var[T]} = \frac{J\,Var(T) + \beta_U \cdot \gamma_U\,Var(U)}{Var(T)} =$$

$$= J + \left(\frac{\beta_U}{\gamma_U}\right) - \text{biased term}$$

causal term

as a solution we can introduce INSTRUMENTAL variable

Matching is performed with the help of some balancing score ⇒ helps find good matches

- **Sensitivity Analysis** – no guarantee that matching leads to balance on variables we didn't match for
  ↳ try to observe the hidden bias

Qs:

#1 Does the conclusion change from statistically significant or not?

#2 Does it change the direction of effect?

Types of Sensitivity Analysis:

① Quick Sanity Check

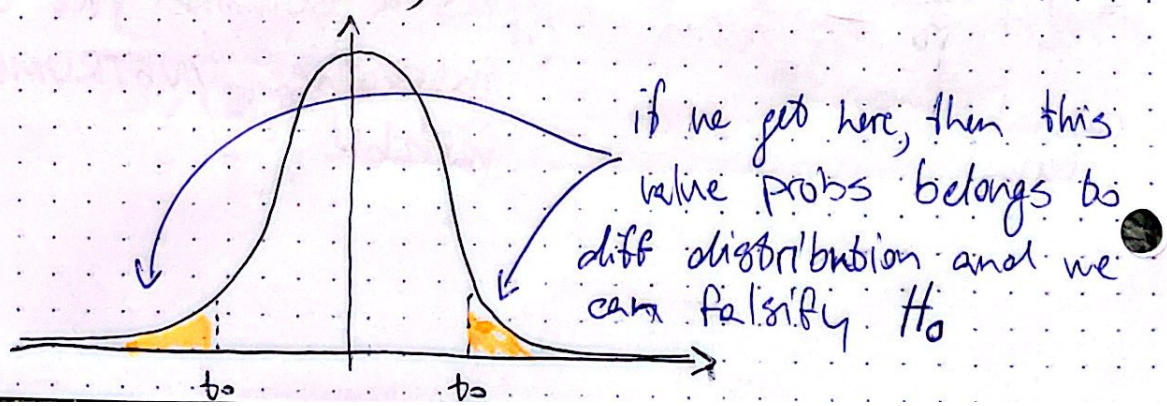
② Super Learning other potential confounders
↳ through cross validation we can see if we are sensitive to certain confounders

③ Deriving bounds on the causal statistical estimation

Hypothesis testing

$$\frac{ATE}{\sigma_{ATE}} = \frac{\text{signal}}{\text{noise}} \sim t\text{-distributed}$$

$$p\text{-value} = Pr\left(\left|\frac{\text{signal}}{\text{noise}}\right| > t_0 \mid H_0\right)$$



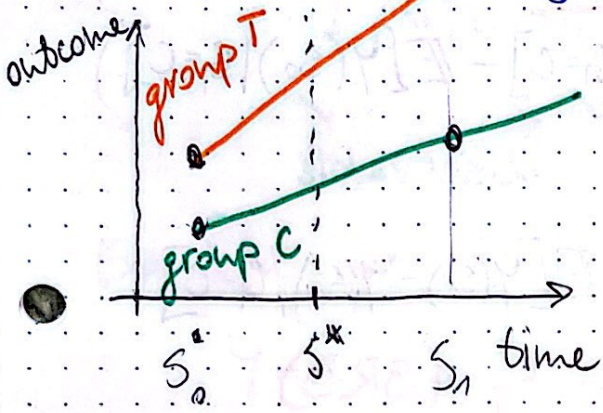
Lecture 8

DiD - Difference in Difference

↳ treatment effect on outcome is estimated as the difference in changes over time between two groups

Group T - receives treatment
Group C - control group

} Both groups receive treatment only after time S^*



! we only have measures at S_1 & S_0 and only know S^* occurred sometime in between

Average Treatment Effect

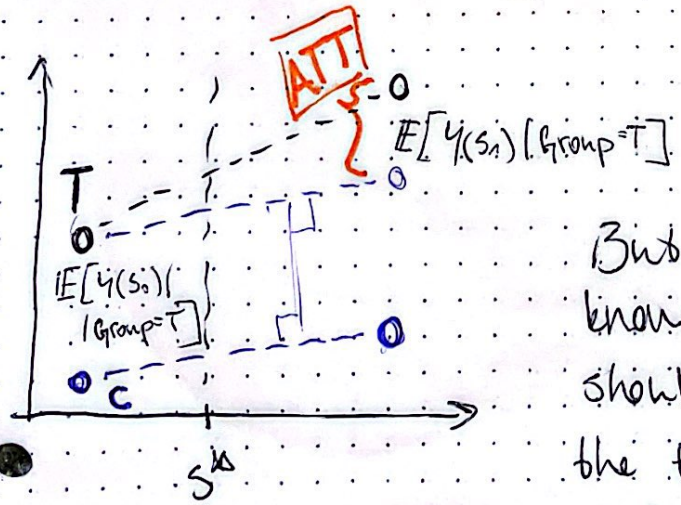
$$\text{ATE} \Rightarrow \mathbb{E}[Y_1 - Y_0] = \mathbb{E}_X\left[\mathbb{E}[Y|T=1,X] - \mathbb{E}[Y|T=0,X]\right]$$

Average Treatment Effect of the Treated

$$\text{ATT} \Rightarrow \mathbb{E}[Y_1 - Y_0 | T=1] = \mathbb{E}_X\left[\mathbb{E}[(Y_1 - Y_0 | T=1, X)]\right] =$$

$$= \mathbb{E}_X\left[\mathbb{E}[Y_1|T=1,X] - \mathbb{E}[Y_0|T=0,X]\right] =$$

$$= \mathbb{E}_X\left[\mathbb{E}[Y|T=1,X] - \mathbb{E}[Y|T=0,X]\right]$$



$$\Rightarrow \text{ATT} = \mathbb{E}[Y_1 - Y_0 | \text{Group} = T]$$

But the problem is that we don't know where T group would score should they not have received the treatment. ~~This is~~ We also need to consider that T & C can be fundamentally different, meaning that

there is a set difference between the two. Now we can approximate that there would a movement across the parallel line from that of c applied to T. The difference between our actual and theoretical points is ATT.

$$ATT = \mathbb{E}[Y_1 - Y_0 \mid G=T] = \left(\mathbb{E}[Y(S_1) \mid G=T] - \mathbb{E}[Y(S_0) \mid G=T]\right) - \left(\mathbb{E}[Y(S_1) \mid G=C] - \mathbb{E}[Y(S_0) \mid G=C]\right)$$

$$\mathbb{E}[Y(S_1) \mid G=T] = \mathbb{E}[Y(S_0) \mid G=T] + \mathbb{E}[Y(S_1) - Y(S_0) \mid G=C]$$

observable

Sharp Regression Discontinuity (SRD)

How can we identify causal effect despite positivity violations?

W determines T

$$T(W) = I\{W \geq c\} = \begin{cases} 1, & W \geq c \\ 0, & W < c \end{cases}$$

c - cutoff value which violates positivity

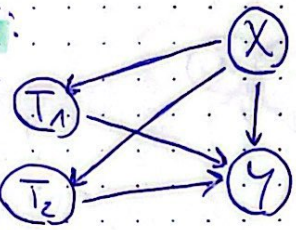
AIM: estimate causal effect: $\tau_{SRD} = \mathbb{E}[Y_1 - Y_0 \mid W=c]$

The SRD looks at discontinuity in outcome at the cut-off
↳ the outcome just above the cut-off minus the outcome just bellow the cut-off

CONDITIONAL AVERAGE TREATMENT EFFECT (CATE)

$$\mathbb{E}[Y_1 - Y_0 | X = x] = ATE$$

Suppose we have 2 treatments:



$$I^a_{i,j} = \Big[\mathbb{E}(Y|(T_i,T_2)=(1,1),X) - \mathbb{E}(Y|(T_1,T_2)=(0,1),X)\Big] - \Big[\mathbb{E}(Y|(T_1,T_2)=(1,0),X) - \mathbb{E}(Y|(T_1,T_2)=(0,0),X)\Big]$$

Example with a linear model:

$$Y = \alpha_0 + \alpha_1 T_1 + \alpha_2 T_2 + \gamma T_1 T_2$$

$$\mathbb{E}(Y|T_1=1,T_2=1) = \alpha_0 + \alpha_1 + \alpha_2 + \gamma$$

$$\mathbb{E}(Y|T_1=1,T_2=0) = \alpha_0 + \alpha_1 \qquad \mathbb{E}(Y|T_1=0,T_2=1) = \alpha_0 + \alpha_2$$

$$\mathbb{E}(Y|T_1=0,T_2=0) = \alpha_0$$

$$ATE_{T_1}(Y|T_2=1) = \alpha_1 + \gamma \qquad ATE_{T_2}(Y|T_1=1) = \alpha_2 + \gamma$$

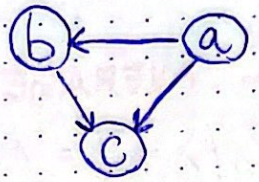
$$ATE_{T_1}(Y|T_2=0) = \alpha_1 \qquad ATE_{T_2}(Y|T_1=0) = \alpha_2$$

$$I^a_{1,2} = \gamma = I^a_{2,1}$$ → Symmetrie!

DAGs

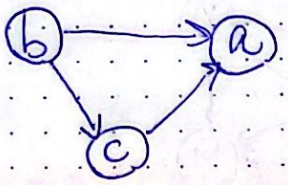
* acyclic!

#1



$$p(a)p(b|a)p(c|a,b) =$$
$$= p(a,b)p(c|a,b) = \underline{p(c,a,b)}$$

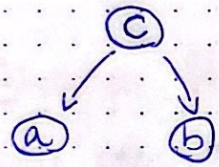
#2



$$p(b)p(c|b)p(a|b,c) =$$
$$= p(b,c)p(a|b,c) = \underline{p(a,b,c)}$$

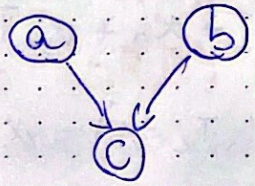
→ we calculate the same probability in different ways

D-separation



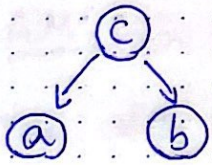
Fork

Chain



Collider

Fork



$$p(a,b,c) = p(c)p(a|c)p(b|c)$$

$$p(a,b) = \sum_c p(a,b,c) = \sum_c p(a|c)p(b|c)p(c) =$$
$$\neq p(a)p(b) \Rightarrow a \not\perp b \mid \emptyset$$

$$p(a,b|c) = \frac{p(a,b,c)}{p(c)} = \frac{p(a|c)p(b|c)p(c)}{p(c)} =$$
$$= p(a|c)p(b|c)$$

$\Rightarrow$ c blocks the path from a to b

$a \perp\!\!\!\perp b \mid c \Rightarrow$ a and b are independent if conditioned on c

Chain

$$p(a,b,c) = p(a)\,p(b|a)\,p(c|b)$$

since summing over all b $\Rightarrow p(c|a)$

$$p(a,b) = \sum_b p(a)p(b|a)p(b|c) = p(a)\underbrace{\sum_b p(b|a)p(b|b)}_{p(c|a)} =$$

$$= p(a)p(c|a) \neq p(a)p(b) \Rightarrow a \not\perp\!\!\!\perp b \mid \emptyset$$

$$p(a,c|b) = \frac{p(a,c,b)}{p(b)} = \frac{p(a)p(b|a)p(c|b)}{p(b)} =$$

bayes:

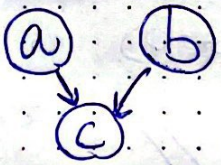
$$= \frac{\cancel{p(a)}p(c|b)}{\cancel{p(b)}} \cdot \frac{p(a|b)\cancel{p(b)}}{\cancel{p(a)}} =$$

$$p(b|a) = \frac{p(a|b)p(b)}{p(a)}$$

$$= p(c|b)\,p(a|b)$$

$\Rightarrow a \perp\!\!\!\perp b \mid c$; c blocks the path from a to b

Collider



$$p(a,b,c) = p(a)p(b)p(c|a,b)$$

$$p(a,b) = \sum_c p(a)p(b)\underbrace{p(c|a,b)}_{=1} = p(a)p(b)\sum_c p(c|a,b) =$$

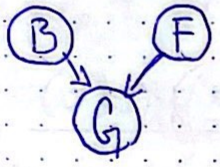
$$= p(a)p(b) \Rightarrow a \perp\!\!\!\perp b \mid \emptyset$$

$$p(a,b|c) = \frac{p(a,b,c)}{p(c)} = \frac{p(a)p(b)p(c|a,b)}{p(c)} \neq p(a)p(b)$$

$$\Rightarrow a \not\perp\!\!\!\perp b \mid \emptyset$$

To conclude, only in case of Collider $a \perp\!\!\!\perp b \mid \emptyset$, for Fork and Chain we know that $a \not\perp\!\!\!\perp b \mid \emptyset$

Fork	Chain	Collider
$a \not\perp\!\!\!\perp b \mid \emptyset$	$a \not\perp\!\!\!\perp b \mid \emptyset$	$a \perp\!\!\!\perp b \mid \emptyset$
$a \perp\!\!\!\perp b \mid c$	$a \perp\!\!\!\perp b \mid c$	$a \not\perp\!\!\!\perp b \mid c$

Exercise: 

$P(B=1) = 0.9 \quad P(F=1) = 0.9$

$P(G=1 \mid B=1, F=1) = 0.8$

$P(G=1 \mid B=1, F=0) = 0.2$

$P(G=0 \mid B=0, F=1) = 0.2$

$P(G=1 \mid B=0, F=0) = 0.1$

$P(F=0) = 1 - P(F=1) = \underline{0.1}$

$$P(F=0 \mid G=0) = \frac{P(F=0, G=0)}{P(G=0)} = \frac{\sum_b P(F=0, B, G=0)}{P(G=0)} =$$

$$= \frac{\left[P(G=0 \mid F=0, B=1) + P(G=0 \mid F=0, B=0)\right] P(F=0, B=0)}{P(G=0)} =$$

$P(F=0)P(B=0)$

$$= \frac{(0.8+0.9) \cdot 0.1 \cdot 0.1}{(0.9+0.2+0.8+0.2)} = //$$

$$= \frac{P(G=0 \mid F=0) P(F=0)}{P(G=0)}$$

$$P(G=0) = \sum_{B,F} P(G=0, B, F) = \sum_{B,F} P(G=0 \mid B, F) P(B \mid F) P(F) =$$

$$= \sum_{B,F} P(G=0 \mid B, F) P(B) P(F) =$$

$$= \#1 \; P(G=0 \mid B=0, F=0) P(B=0) P(F=0) +$$

$$\#2 + P(G=0 \mid B=1, F=0) P(B=1) P(F=0) +$$

$$\#3 + P(G=0 \mid B=1, F=1) P(B=1) P(F=1) +$$

$$\#4 + P(G=0 \mid B=0, F=1) P(B=0) P(F=1) =$$

$[0.9 \cdot 0.1 \cdot 0.1] + [0.8 \cdot 0.9 \cdot 0.1] + [0.2 \cdot 0.9 \cdot 0.9] + [0.2 \cdot 0.1 \cdot 0.9] =$

$= 0.315$ ✓

$$P(G=0 \mid F=0) = \sum_B P(G=0 \mid F=0, B) P(B) =$$

$$= P(G=0 \mid F=0, B=0) P(B=0) + P(G=0 \mid F=0, B=1) P(B=1) =$$

$$= 0.9 \cdot 0.1 + 0.8 \cdot 0.9 = \underline{0.81}$$ ✓

$$= \frac{P(G=0 \mid F=0) P(F=0)}{P(G=0)} = \frac{0.315 \cdot 0.1}{0.81} = 0.0389$$

---

path p is blocked by a set of nodes Z iff:

1) p contains chain $A \to B \to C$ or a fork $A \leftarrow B \to C$ such that B is in Z

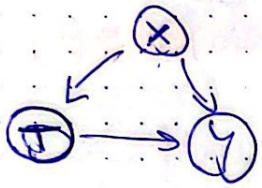
2) p contains collider $A \to B \leftarrow C$ such that collision node B is not in Z and no descendants of B is in Z

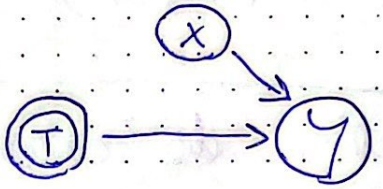
Lecture 10

Observation (conditioning) vs Intervention

Learn to distinguish between: var T taking a value t naturally and cases where we 'fix' $T=t$ by denoting the letter $do(T=t)$

$$p(Y=y \mid T=t) \qquad p(Y=y \mid do(T=t))$$





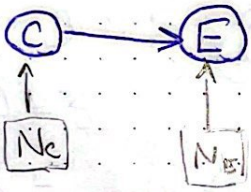
- probability of $Y=y$ when we 'intervene' to make $T=t$

#graphSergery

Example:

$C := N_C$

$E := 4 \cdot C + N_E$



$N_C, N_E \sim \mathcal{N}(0,1)$

apply $do(C)$

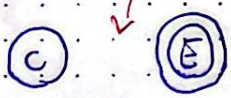
$$p(E \mid do(C)) \neq p(E)$$

$$= p(E \mid C)$$

since there are no other confounders

apply $do(E)$

we broke the arrow



$$p(C \mid do(E)) = p(C)$$

$$\neq p(C \mid E)$$

hence we get this!

Intervention example $C := N_C \quad E := 4 \cdot C + N_E$

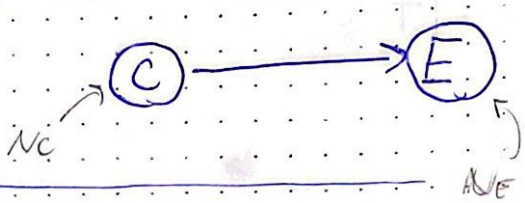
$N_C, N_E \sim \mathcal{N}(0,1), \; N_C \perp\!\!\!\perp N_E$

$Var[aX] = a^2 Var[X], \; 4C \sim \mathcal{N}(0,16)$

And since $4C \perp\!\!\!\perp N_E$, and the sum of two normally distributed random vars:

$E \sim \mathcal{N}(\mu_{4C} + \mu_{N_E}, \sigma_{4C}^2 + \sigma_{N_E}^2)$

$E \sim \mathcal{N}(0, 17)$



Now let's **do**-calculus

$do(C=2) \Rightarrow p(E \mid do(C=2))$: $C = 2$

$E = 4 \cdot C + N_E =$

$= 8 + N_E$

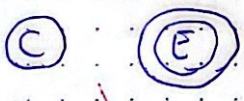
$N_E \sim \mathcal{N}(0,1)$

$E \sim \mathcal{N}(\underset{8}{\mu_{4C}} + \underset{0}{\mu_{N_E}}, \underset{0}{\sigma_{4C}^2} + \underset{1}{\sigma_{N_E}^2})$

$E \sim \mathcal{N}(8, 1)$

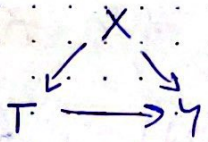
$= p(E \mid C=2)$

$p(C \mid do(E=2)) = p(C) = \mathcal{N}(0,1)$



cuz we break the relation (=bond)

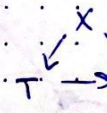
Adjustment Formula

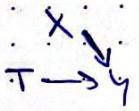


T - drug usage
X - sex
Y - recovery

Estimate effect of treatment:

$P(Y=1 \mid do(T=1)) - P(Y=1 \mid do(T=0))$

$P(Y=y \mid do(T=b))$ for  is equal to

$P_m(Y=y \mid T=b)$ for modified graph 

(modified! (do calculus was applied))

$$P(Y=y \mid do(T=b)) = P_m(Y=y \mid T=b) =$$

$$= \sum_x P_m(Y=y \mid T=b, X=x) \cdot P_m(X=x \mid T=b) =$$

$$= \sum_x P_m(Y=y \mid T=b, X=x) P_m(X=x)$$

since T and X aren't connected

Exercise:	Drug		No drug		T - drug
Men	81/87	(93%)	234/270	(87%)	X - sex
Women	192/263	(73%)	55/80	(69%)	Y - outcome
Total:	273/350	(78%)	289/350	(83%)	

$ATE = ACE = P(Y=1 \mid do(T=1)) - P(Y=1 \mid do(T=0)) = ?$

$$P(Y=1 \mid do(T=1)) = \sum_x P_m(Y=1 \mid T=1, X=x) P(X=x) =$$

$$= P_m(Y=1 \mid T=1, X=1) P(X=1) + P_m(Y=1 \mid T=1, X=0) P(X=0) =$$

$$= 0.93 \cdot (87+270)/700 + 0.73 \cdot (263+80)/700 =$$

$$= 0.832$$

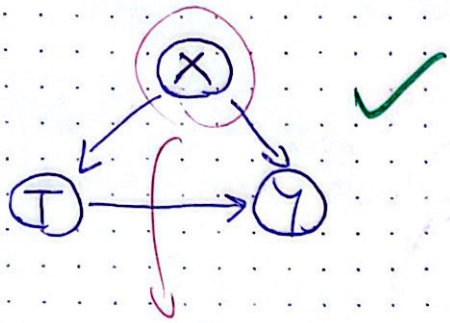
$$P(Y=1 \mid do(T=0)) = P_m(Y=1 \mid T=0, X=0) P(X=0) + P_m(Y=1 \mid T=0, X=1) P(X=1) =$$

$$= 0.69 (263+80)/700 + 0.87 (87+270)/700 = 0.7818$$

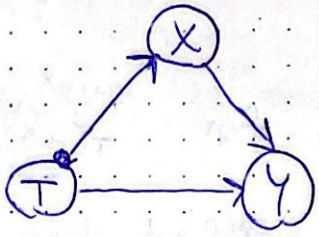
$$0.832 - 0.7818 = \boxed{0.0505}$$

When to use this Backdoor approach?

$$P(Y=y \mid do(T=t)) = \sum_x P_m(Y=y \mid T=t, X)\, P_m(X)$$

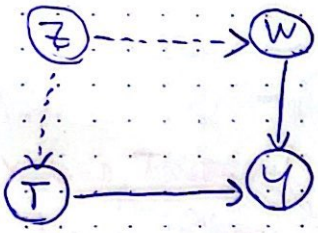


✓



✗

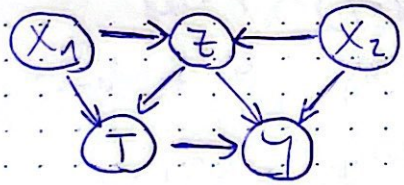
here X (= confounder) acts as a backdoor



We can use W as a backdoor!

1) W blocks $T \leftarrow Z \rightarrow W \rightarrow Y$
2) W leaves directed p $T \rightarrow Y$ uninterrupted
3) W is not a collider or descendants of T

$$P(Y=y \mid do(T=t)) = \sum_w P_m(Y=y \mid T=t, W=w)\, P_m(W=w)$$



can we condition (use as backdoor) on Z? It's a collider, which is a problem. However, if we condition on ZX_1 or ZX_2 or ZX_1X_2 then we overrule and it's okay!

Lecture 11

Total Law of Probability:

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

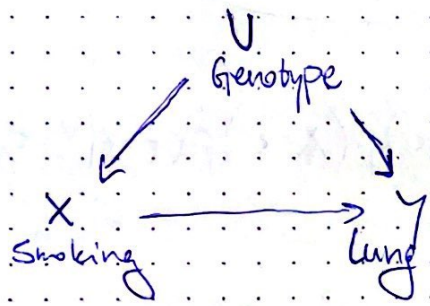
$$P(X=1|Y=1)P(Y=1) + P(X=1|Y=0)P(Y=0) = P(X=1)$$

$X \rightarrow Y$ $\quad P(X|Y) \neq P(X)$

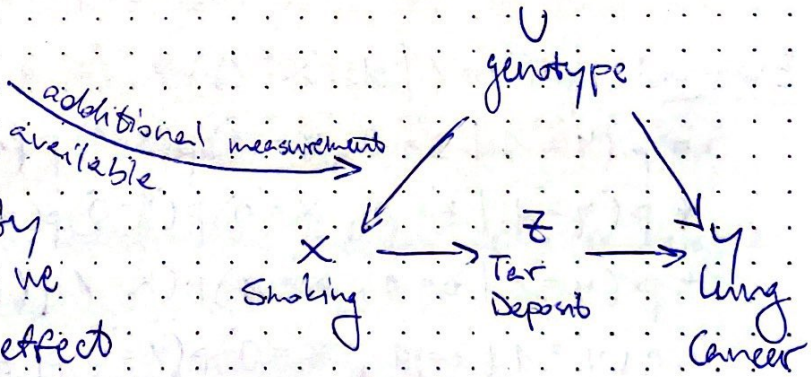
$X \rightarrow Y \rightarrow W$ $\quad P(X|W) \neq P(X)$

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

Pearl's Front-Door Adjustment



Graph does not satisfy backdoor criteria since quantity we need to condition on to block the path (~~lung cancer~~ genotype) is unobserved



Graph still doesn't satisfy the backdoor criteria but we can determine the causal effect

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

1) $X \rightarrow Z$ is identifiable, since no back path from X and Z.

$$P(Z=z|do(X=x)) = P(Z=z|X=x)$$

2) $Z \rightarrow Y$ is identifiable, since backdoor from Z to Y:

$$Z \leftarrow X \rightarrow U \rightarrow Y$$

is blocked by conditioning on X:

$$p(Y=y \mid do(Z=z)) = \sum_x p(Y=y \mid Z=z, X=x)\, p(X=x)$$

By letting z be the value Z takes when setting $X=x$:

$$p(Y \mid do(X)) = p(Y \mid do(X), Z) = p(Y \mid do(Z=z))$$

$$\downarrow$$

$$p(Y=y \mid do(X=x)) = \sum_z p(Y=y, z \mid do(X=x))$$

$$= \sum_z p(Y=y \mid z, do(X=x))\, p(z \mid do(X=x))$$

$$= \sum_z p(Y=y \mid do(Z=z))\, p(z \mid do(X=x))$$

$$p(Y=y \mid do(X=x)) = \sum_z \sum_{x'} p(Y=y \mid z=z, X=x')\, p(X=x')\, p(Z=z \mid X=x)$$

Example: $p(Y=1 \mid do(X=1)) =$

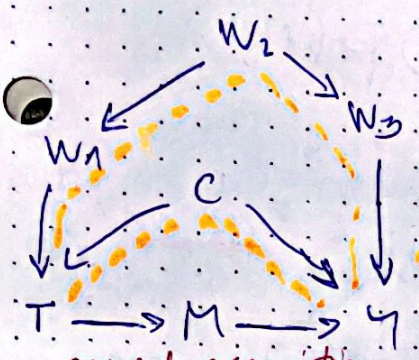
$= p(Y=1 \mid z=0, X=1)\, p(X=1)\, p(z=0 \mid X=1)$

$+ p(Y=1 \mid z=0, X=0)\, p(X=0)\, p(z=0 \mid X=1)$

$+ p(Y=1 \mid z=1, X=1)\, p(X=1)\, p(z=1 \mid X=1)$

$+ p(Y=1 \mid z=1, X=0)\, p(X=0)\, p(z=1 \mid X=1) =$

Backdoor Adjustment – we want to block all paths from T to Y that can act as backdoors



causal association flow

not a backdoor path!

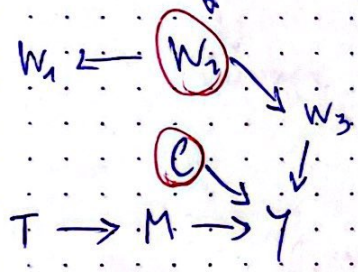
$P(Y|T=t)$

backdoor paths = non-causal association paths

⇓

we want to block these and keep only causal association paths

Intervene $do(T=t)$



* we can remove all incoming flows to T

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

we can calculate this by conditioning on C and W_2 from first graph! this then blocks the backdoor

CRITERION

A set of variables W satisfies the backdoor criterion relative to T and Y if:

1) W blocks all backdoor paths from T to Y
2) W does not contain any descendants of T

Then we can identify causal effects of T on Y:

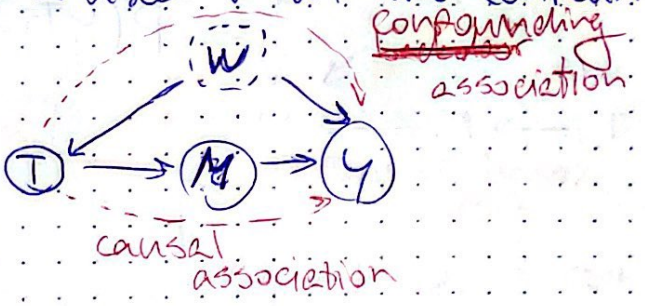
$$P(Y|do(T=t)) = \sum_w P(Y|t,w)P(w)$$

ADJUSTMENT

for our example that would give us:

$$P(y \mid do(T=t)) = \sum_w \sum_c p(y \mid t, w, c)\, p(w, c)$$

FRONT DOOR ADJUSTMENT – used when the confounder is unobservable



1) Identify causal effect of T on M
2) Identify causal effect of M on Y
3) Combine 1) & 2) to get causal effect of T on Y

1 $p(m \mid do(t)) = p(m \mid t)$

no backdoor paths from T to M.
↳ cuz Y is a collider

backdoor adjustment

2 $p(y \mid do(m)) = \sum_t p(y \mid t, m)\, p(t)$

there is a backdoor path from M to Y
$M \leftarrow T \leftarrow W \rightarrow Y$
} we can condition on T

3 $p(y \mid do(t)) = \sum_m p(m \mid do(t))\, p(y \mid do(m)) =$

ADJUSTMENT

$= \sum_m p(m \mid t) \sum_{t'} p(y \mid t', m)\, p(t')$

chaining them together and summing over middle var ($= m$)

CRITERION

1) M completely mediates effect of T on Y (causal paths from T to Y go through M)
2) There is no unblocked path from T to M
3) All backdoor paths from M to Y are blocked by T

Lecture 12

MARKOV BLANKET AND BOUNDRY

Markov blanket of random variable

- Y in a random variable set $S = \{X_1, \dots, X_n\}$ is any subset S_1 of S, conditioned on which other variables are independent with Y: $Y \perp\!\!\!\perp S \setminus S_1 \mid\mid S_1$

It means S_1 contains all the information needed to infer Y and the variables $S \setminus S_1$ are redundant

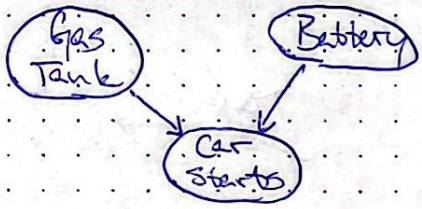
⇒ In general Markov blanket is not unique. Any set S that contains Markov blanket is a Markov blanket itself

Markov Boundry is a Markov blanket none of whose subsets are Markov blankets themselves

DISTINGUISHING CAUSAL STRUCTURES: V-structures

Gas tank $\perp\!\!\!\perp$ battery

Gas tank $\not\perp\!\!\!\perp$ battery | car starts = 0



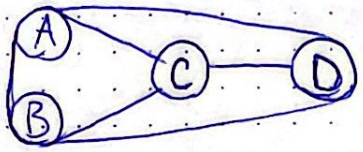
- **Markov Equivalence Class (MEC)** – graphs G and G' belong to the same equivalence class iff each conditional independence implied by G is also implied by G' and vice versa.

$A \perp\!\!\!\perp C \mid B$:

1) $A \to B \to C$
2) $A \to B \leftarrow C$
3) $A \leftarrow B \leftarrow C$

Peter-Clark (PC) Algorithm

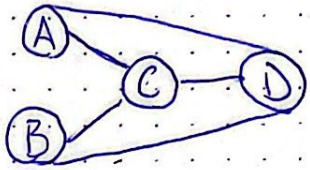
1) Start with complete graph:



2) Zeroth order Conditional Independence; $A \perp\!\!\!\perp B$

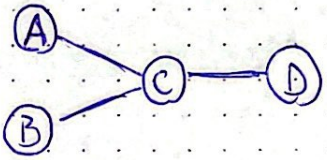
we perform statistical tests to see if they are independent or not

if not "||"

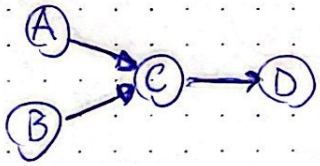


3) First order CI; $A \perp\!\!\!\perp D \mid C$ & $B \perp\!\!\!\perp D \mid C$:

4) No higher CI order observed :)



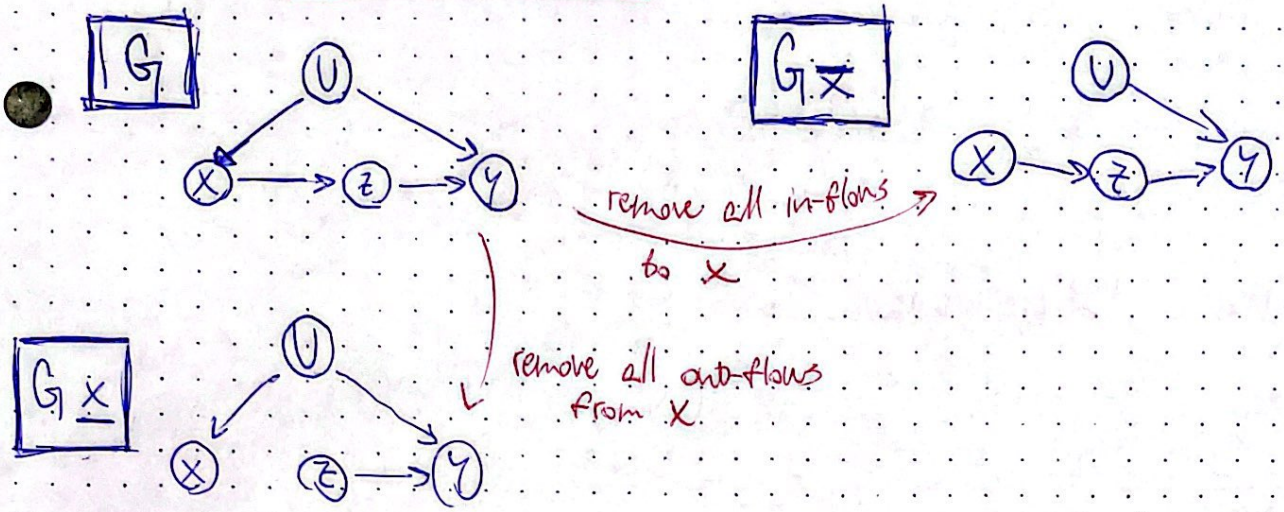
5) Sometimes we can't put the arrows, but here we have a collider:



lecture 13

Lecture 13

DO-CALCULUS



Rule 1: (insertion/deletion of observations)

$$P(Y \mid do(X=x), Z, W) = P(Y \mid do(X=x), W) \text{ if } (Y \perp\!\!\!\perp Z) \mid X, W \text{ in } G_{\bar{X}}$$

↳ if Y and Z are d-separated by X, W in a graph where incoming edges in X have been removed

---

d-separation – two sets of nodes X and Y are d-separated by a set of nodes Z if all of the paths between X and Y are blocked by Z.

Rule 2: (action/observation exchange)

$$P(Y \mid do(X=x), do(Z=z), W) = P(Y \mid do(X=x), z, W) \text{ if } (Y \perp\!\!\!\perp Z) \mid X, W \text{ in } G_{\bar{X}\underline{Z}}$$

↳ if Y and Z are d-separated by X, W in a graph where incoming edges in X and outgoing edges from Z have been removed.

Rule 3. Insertion/deletion of actions

$p(y \mid do(x), do(z), w) = p(y \mid do(x), w)$ if $Y \perp\!\!\!\perp Z \mid X, W$ in $G_{\overline{XZ(W)}}$

$Z(W)$ is set of Z-nodes that are not ancestors of any W-node in $G_{\overline{X}}$

Optimal Adjustment Sets

$$pa_G(en_G(X \to Y)) \setminus (en_G(X \to Y) \cup \{X\})$$

tells us to condition on all parents of Y excluding those on path $X \to Y$ $\Rightarrow$ THEOREM

LECTURE 14

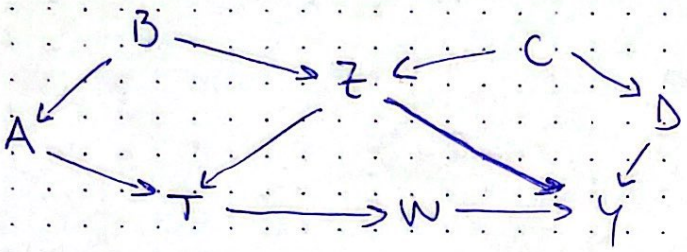
$do(T = g(z))$ where: $g(z) = \begin{cases} 1 & \text{when } Z > z \\ 0 & \text{otherwise} \end{cases}$

result of such policy is: $p(Y = y \mid do(T = g(z)) = ?$

1) Introduce "z specific effect" of T on Y

$$p(Y = y \mid do(T = t), Z = z) = \sum_s p(Y = y \mid T = t, S = s, Z = z) P(S = s \mid Z = z)$$

Lecture 14



We need to condition on Z, which is a collider

↓

need to block the spurious path by condition on A (parent of T)

Q: What's the causal effect of T on Y?

$$p(Y=y \mid do(T=t)) =$$

$$= \sum_{z,a} p(Y=y \mid T=t, A=a, Z=z)\, p(Z=z, A=a)$$

BACKDOOR ADJUSTMENT

Notes: Collider → closed path —condition→ ⇒ opened

Other → opened path ⇒ closed

So here we must obviously condition on Z, but this is not enough cuz it gets OPENED. Hence, must condition on another node along that path ⇒ (A, B, C, D)

Q: What's c-specific effect of T on Y?

↳ need to condition on C!

$$p(Y=y \mid do(T=t), C=c) = \sum_{z} p(Y=y \mid T=t, C=c, Z=z)\, p(Z=z \mid C=c)$$

Q: What's z-specific effect of T on Y?

$$p(Y=y \mid do(T=t), Z=z) = \sum_{c} p(Y=y \mid T=t, Z=z, C=c)\, p(C=c \mid Z=z)$$

Q: What is z-dependent causal effect T on Y under strategy: $g(z) = \begin{cases} 0 & ; z \leq 2 \\ 1 & ; z > 2 \end{cases}$; $z \in \{1,2,3,4,5\}$

$$p(Y=y \mid do(T=g(z))) = \sum_z p(Y=y \mid do(T=g(z)), Z=z)\, p(Z=z) =$$

$$= p(Y=y \mid do(T=0), Z=1)\, p(Z=1)$$
$$+ p(Y=y \mid do(T=0), Z=2)\, p(Z=2)$$
$$+ p(Y=y \mid do(T=1), Z=3)\, p(Z=3)$$
$$+ p(Y=y \mid do(T=1), Z=4)\, p(Z=4)$$
$$+ p(Y=y \mid do(T=1), Z=5)\, p(Z=5) \quad *$$

⇓ we can further expand by using result from z-specific effect (prev. page)

$$* \left(p(Y=y \mid do(T=t), Z=z) = \sum_c p(Y=y \mid T=t, Z=z, C=c)\, p(C=c \mid Z=z) \right)$$

$$\Rightarrow \sum_c p(Y=y \mid T=0, Z=1)\, p(Z=1)\, p(C=c \mid Z=1) +$$
$$+ \sum_c p(Y=y \mid T=0, Z=2, C=c)\, p(Z=2)\, p(C=c \mid Z=2) +$$
$$+ \sum_c p(Y=y \mid T=1, Z=3, C=c)\, p(Z=3)\, p(C=c \mid Z=3) +$$
$$+ \sum_c p(Y=y \mid T=1, Z=4, C=c)\, p(Z=4)\, p(C=c \mid Z=4) +$$
$$+ \sum_c p(Y=y \mid T=1, Z=5, C=c)\, p(Z=5)\, p(C=c \mid Z=5)$$