

Causality

Talk 3: Causal DAG Models (Causal Bayesian networks)

Yao Zhang

Intelligent Information Processing Research Group,
Faculty of Electrical Engineering and Computer Science,
Ningbo University

Note: The following slides are primarily adapted from the course materials¹.

Nov 28, 2025

¹C. Heinze-Deml. Causality. URL: <https://stat.ethz.ch/lectures/ss21/causality.php>.

Task

Causal inference-based attention model

Attentional modeling is a significant concept in neural networks, with a principle similar to human vision. The model focuses its attention on a specific part, disregarding information in other locations. By filtering out extraneous visual information, the attention model increases visual recognition efficiency, enabling us to better comprehend images, language, and other data. This approach has also been widely applied to neural networks, with a significant impact.

Wang et al. [207] introduced a causal attention module (CaaM) that self-annotates confounders in an unsupervised manner, using causal intervention to eliminate the effects of confounders. The article hypothesizes that attention has opposite effects in IID and OOD tasks, and that attentional modeling is less effective than nonattentional baseline modeling in OOD tasks. The authors attribute this to confounding effects. To eliminate the impact of these confounding factors, they propose a causal intervention approach. Specifically, they constructed a causal graph to describe the relationships between the input image X , the label Y , the confounder S , and the mediating variable M . Using a data partitioning intervention method, the training data $T = \{t_1, \dots, t_m\}$ are partitioned, with each partition representing a confounding layer. A back-door adjustment is employed to cut off the back-door path $X \leftarrow S \rightarrow Y$. Intervention is performed through data partitioning and iterative self-annotation of confounders. To avoid overfitting, adversarial learning is used to separately learn causal and confounding features.

Figure 1: Causal inference-based attention model in the article².

²L. Jiao et al. “Causal Inference Meets Deep Learning: A Comprehensive Survey”. In: *«Research»* 7 (2024), pp. 1–41.   

Factorization of the joint density

- We always have:

$$f(x_1, \dots, x_p) = f(x_1)f(x_2|x_1) \dots f(x_p|x_1, \dots, x_{p-1})$$

- A set of variables $X_{\text{pa}(j)}$ is said to be **Markovian parents** of X_j if it is a minimal subset of $\{X_1, \dots, X_{j-1}\}$ such that $f(x_j|x_1, \dots, x_{j-1}) = f(x_j|x_{\text{pa}(j)})$
- Then

$$f(x_1, \dots, x_p) = \prod_{j=1}^p f(x_j|x_{\text{pa}(j)})$$

"factorization property"

- We can draw a DAG accordingly
- The distribution is said to **factorize** according to this DAG

Last time, Cont.

- First-order Markov models: the future is independent of the past given the present

$$1 \rightarrow 2 \rightarrow \cdots \rightarrow t-1 \rightarrow t \rightarrow t+1$$

$$X_{t+1} \perp\!\!\!\perp \{X_1, X_2, \dots, X_{t-1}\} \mid X_t$$

- In DAG models, we have a similar (local) Markov property
- Let S be any collection of nodes. Then:

$$X_S \perp\!\!\!\perp X_{V \setminus \text{desc}(S) \setminus \text{pa}(S)} \mid X_{\text{pa}(S)} \quad (1)$$

When there are three random variables forming a collider structure (i.e., $X \rightarrow Z \leftarrow Y$), $X \perp\!\!\!\perp Y \mid Z$?

In more general cases, we need d-separation.

d-separation

- A path between i to j is blocked by a set S (not containing i or j) if at least one of the following holds:
 - There is a non-collider on the path that is in S ; or
 - There is a collider on the path such that neither this collider nor any descendants are in S
- A path that is not blocked is active
- If all paths between $i \in A$ and $j \in B$ are blocked by S , then A and B are d-separated by S . Otherwise they are d-connected given S .
- Denote d-separation by $\perp\!\!\!\perp$

Global Markov property

- **Definition:**

A distribution P with density p satisfies the **global Markov property** with respect to a DAG G if:

$$A \text{ and } B \text{ are d-separated by } S \text{ in } G \Rightarrow X_A \perp\!\!\!\perp X_B \mid X_S \text{ in } P$$

- **Theorem (Pearl, 1988):**

A distribution P with density p satisfies the global Markov property with respect to G if and only if p factorizes according to G .



Causal graphical models

Causality

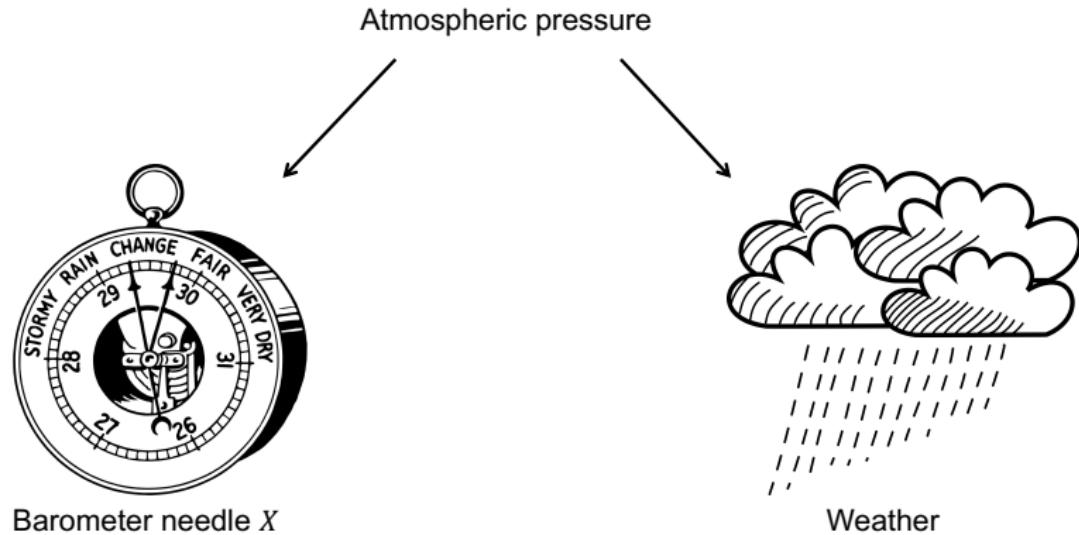
Christina Heinze-Deml

Spring 2021

Today

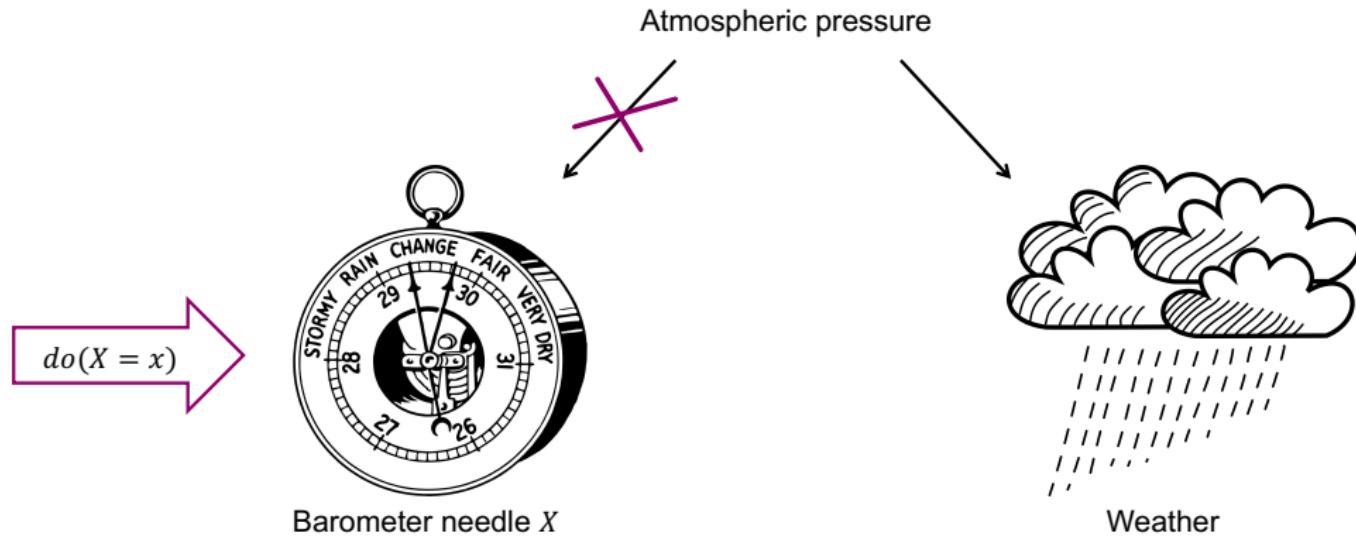
- Causal effects and do-operator
- Causal graphical models
- Selection bias

Causal effect – Example



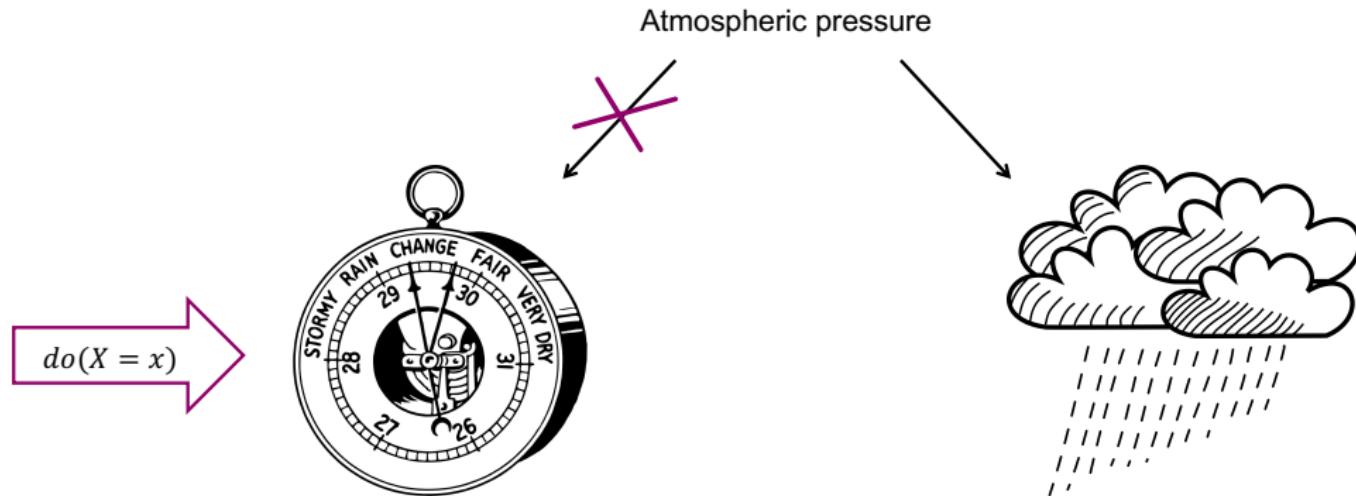
Example due to Frederick Eberhardt

Causal effect – Example



Example due to Frederick Eberhardt

Causal effect – Example



X has a **causal effect** on Y if manipulating X changes the distribution of Y

Example due to Frederick Eberhardt

Causal effect and do-operator

- Interventional definition of causal effect:
X has a causal effect on Y if manipulating X changes the distribution of Y
- Mathematical notion of manipulation (due to Pearl):
 - $do(X = x)$ (or shorthand $do(x)$) represents a hypothetical intervention where X is set to the value x , uniformly over the entire population
 - $p(y|do(X = x))$ is the distribution of Y after $do(X = x)$
 - $E(Y|do(X = x))$ is the expectation of Y after $do(X = x)$, etc

Conditioning on observing: $p(y|see(X = x)) = p(y|x)$ (ordinary conditioning)

Intervening: $p(y|do(X = x))$, also written as $p^{do(X=x)}(y)$

Causal effect and do-operator

- Mathematical definition of causal effect:

X has a causal effect on Y if $p(y|do(X = x'))$ depends on x' ,
 i.e., if $\exists a$ and b so that $p(y|do(X = a)) \neq p(y|do(X = b))$

- Total average causal effect:

$$\text{ACE}(x, x') = \underbrace{E(y|do(X = x))}_{\uparrow} - \underbrace{E(y|do(X = x'))}_{\uparrow}$$

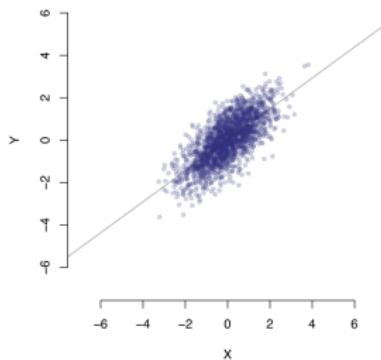
Eg X binary, treatment $X = 1$
 no treatment $X = 0$

$$\text{ACE} = \underbrace{E(y|do(x=1))}_{\sim\sim} - \underbrace{E(y|do(x=0))}_{\sim\sim}$$

Classical regression models

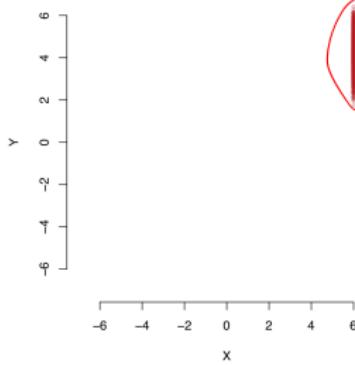
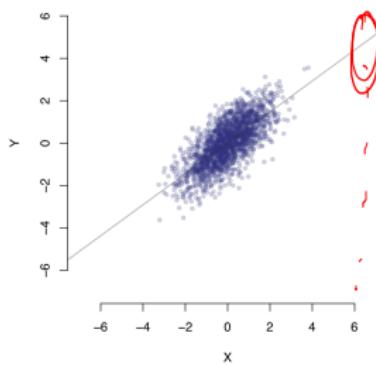
- We observe n i.i.d. observations of (X, Y)
- Goal is to model certain aspects of $p(y|x)$, for example $E(Y|X = x)$
- Useful for prediction

Model aspects of distribution of Y when we observe $X = x$

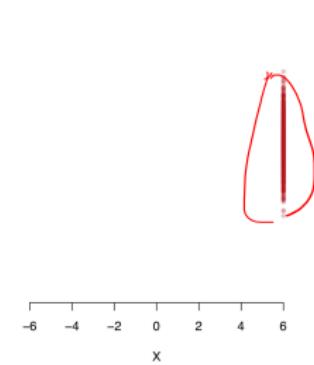


Classical regression models

- We observe n i.i.d. observations of (X, Y)
- Goal is to model certain aspects of $p(y|x)$, for example $E(Y|X = x)$
- Useful for prediction – but what if we set X to e.g. 6? I.e. if $do(X = 6)$? 



$X \rightarrow Y$
 X



$Y \rightarrow X$
 Y

Classical regression models

- We **observe** n i.i.d. observations of (X, Y)
- Goal is to model certain aspects of $p(y|x)$, for example $E(Y|X = x)$
- Useful for prediction
- Such analyses are generally **not** useful for policy or treatment decisions, since such decisions involve predictions in manipulated systems with **post-intervention distributions** different from p

Example

- Consider a rehabilitation program for prisoners. Participation in the program is voluntary.
 - $X = 1$ if prisoner participated in the program; $X = 0$ otherwise
 - $Y = 1$ if prisoner is rearrested within a year; $Y = 0$ otherwise
- $P(Y = 1|X = 1)$: probability of re-arrest for prisoners who choose to participate
- $P(Y = 1|do(X = 1))$: probability of re-arrest if program were compulsory for all prisoners
- Note that generally $P(Y = 1|do(X = 1)) \neq P(Y = 1|X = 1)$



Example

- Suppose $P(Y = 1|X = 1) < P(Y = 1|X = 0)$
 - Re-arrest rate among prisoners who participated in the program is lower than among those who did not participate
 - Could be due to the program, due to the intrinsic motivation of the prisoners who chose to participate, due to a mixture of these two, or....
- Suppose $P(Y = 1|do(X = 1)) < P(Y = 1|do(X = 0))$
 - Program lowers the re-arrest rate, i.e., program has a causal effect on the re-arrest rate
 - Manipulating X changes the distribution of Y
 - X is causal for Y

Frameworks

- Causal DAG models (Causal Bayesian networks)
- Structural equation models
- Potential outcomes

Causal Bayesian networks

- Let $G = (V, E)$ be a DAG and P be the distribution of X_V with density p
- The pair (G, P) is a **DAG model** or a **Bayesian network** if

$$p(x_V) = \prod_{i \in V} p(x_i | x_{\text{pa}(i)})$$

Causal Bayesian networks

- Let $G = (V, E)$ be a DAG and P be the distribution of X_V with density p
- The pair (G, P) is a **DAG model** or a **Bayesian network** if

$$p(x_V) = \prod_{i \in V} p(x_i | x_{\text{pa}(i)})$$

- The pair (G, P) is a **causal DAG model** or a **causal Bayesian network** if for any $W \subset V$

$$p(x_V | do(x_W = x'_W)) = \prod_{i \in V \setminus W} p(x_i | x_{\text{pa}(i)}) \mathbf{1}\{x_W = x'_W\}$$

Causal Bayesian networks

- The pair (G, P) is a causal DAG model or a causal Bayesian network if for any $W \subset V$

$$p(x_V | do(x_W = x'_W)) = \prod_{i \in V \setminus W} p(x_i | x_{\text{pa}(i)}) \mathbf{1}\{x_W = x'_W\}$$

- Modified factorization known as
 - “g-formula” (Robins)
 - “manipulation formula” (Spirtes, Glymour, Scheines)
 - “truncated factorization formula” (Pearl)

Causal Bayesian networks

- The truncated factorization formula implies that an intervention on X_j only changes $p(x_j | x_{\text{pa}(j)})$; the other conditional distributions remain unchanged. This is also known as **invariance**.

Compare: $p(x_v) = \left\{ \prod_{i \in V \setminus \{j\}} p(x_i | x_{\text{pa}(i)}) \right\} \underbrace{p(x_j | x_{\text{pa}(j)})}_{\text{orange}}$

$$p(x_v | \text{do}(x_j = x'_j)) = \left\{ \prod_{i \in V \setminus \{j\}} p(x_i | x_{\text{pa}(i)}) \right\} \cdot \underbrace{\prod_{h \in \text{pa}(j)} h x_h = x'_h}_{\text{orange}}$$

Notes week 3 - I

A : Altitude

T : Temperature

$$A \rightarrow T$$

$p(a, t)$: joint density of the altitude A and the average temperature T

$$p(a, t) = \underbrace{p(t|a)}_{\text{invariant physical mechanism}} p(a)$$

→ if we change the altitude A, then we assume that the physical mechanism $p(t|a)$ responsible for producing an average temperature is still in place (invariance)

→ holds independent of $p(a)$

Causal Bayesian networks

- The truncated factorization formula implies that an intervention on X_j only changes $p(x_j | x_{\text{pa}(j)})$; the other conditional distributions remain unchanged. This is also known as **invariance**.

"Interventions act locally"

But note: can have interventions on multiple nodes at the same time

$$p(x_v | do(x_j = x'_j), do(x_e = x'_e)) = \prod_{i \in v \setminus \{j, e\}} p(x_i | x_{\text{pa}(i)}) \cdot \mathbb{1}_{\{x_j = x'_j\}} \cdot \mathbb{1}_{\{x_e = x'_e\}}$$

Causal Bayesian networks

- The pair (G, P) is a causal DAG model or a causal Bayesian network if for any $W \subset V$

$$p(x_V | do(x_W = x'_W)) = \prod_{i \in V \setminus W} p(x_i | x_{\text{pa}(i)}) \underbrace{1\{x_W = x'_W\}}$$

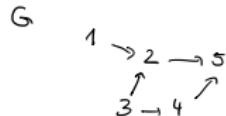
post-intervention distributions
needed to define causal
effects

conditional distribution that
can be estimated from
observational data

Causal Bayesian networks

- The modified factorizations represent factorizations wrt truncated graphs G_W , where all edges into W are removed
- See Notes week 3 – II.pdf

Notes week 3 - II



$$p(x_1, x_2, x_3, x_4, x_5) = p(x_1)p(x_3)p(x_2|x_1, x_3)p(x_4|x_3)p(x_5|x_2, x_4)$$

Consider a do-intervention on X_2 :

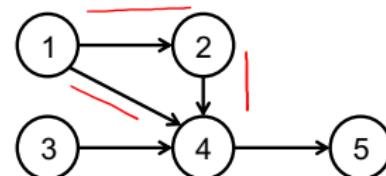


$$p(x_1, x_2, x_3, x_4, x_5 | \text{do}(x_2 = x'_2)) = p(x_1)p(x_3) \cdot \underbrace{p(x_2 | x'_2)}_{p(x_2 | x_3)} \cdot \\ p(x_4 | x_3)p(x_5 | x_2, x_4)$$

$$p(x_1, x_3, x_4, x_5 | \text{do}(x_2 = x'_2)) = p(x_1)p(x_3)p(x_4|x_3)p(x_5|x_2, x_4) \Big|_{x_2=x'_2}$$

Causal Bayesian networks

- $X_{\text{pa}(j)}$ can be interpreted as the direct causes of X_j
- Directed edges can be interpreted as **direct causal effects**
- X can only be causal for Y if there is a directed path from X to Y
- Depending on the context, we might be interested in the
 - Direct causal effect
 - Indirect causal effect
 - Total causal effect
 - ...



Notes week 3 - III

Consider the DAGs $1 \rightarrow 2$ and $2 \rightarrow 1$

- * Any distribution P of (x_1, x_2) factorizes wrt these two DAGs:

$$p(x_1, x_2) = p(x_1) p(x_2 | x_1) = p(x_2) p(x_1 | x_2)$$

Assume $x_1 \perp\!\!\!\perp x_2$

- * But the DAGs are very different when interpreted causally:

Consider a do-intervention on x_2 .

The post-intervention distribution of x_1 is:

- 1) For the causal DAG $1 \rightarrow 2$

$$p(x_1 | \text{do}(x_2 = x'_2)) = p(x_1) \Big|_{\substack{x_2 = x'_2 \\ \text{does not depend on } x_2 \\ (\text{no causal effect from } x_2 \text{ to } x_1)}}$$

$$p(x_1 | \text{do}(x_2 = x'_2)) = p(x_1 | \text{do}(x_2 = \tilde{x}_2)) = p(x_1)$$

- 2) For the causal DAG $2 \rightarrow 1$

$$\begin{aligned} p(x_1 | \text{do}(x_2 = x'_2)) &= p(x_1 | x_2) \Big|_{x_2 = x'_2} \\ &= p(x_1 | x'_2) \quad \text{Causal effect} \\ &\quad \text{of } x_2 \text{ on } x_1 \end{aligned}$$

Here, $p(x_1 | \text{do}(x_2 = x'_2)) \neq p(x_1 | \text{do}(x_2 = \tilde{x}_2))$, $x'_2 \neq \tilde{x}_2$

Simpson's paradox

	Treatment	Placebo
Male	50/100	150/500
Female	50/500	0/100
Total	100/600	150/600

↓ replace gender by blood pressure (BP); numbers stay the same

	Treatment	Placebo
High BP	50/100	150/500
Low BP	50/500	0/100
Total	100/600	150/600

Simpson (1951), in an example similar to this one:
"The treatment can hardly be rejected as valueless to the race when it is beneficial when applied to males and to females."

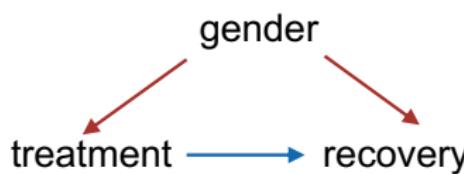
⇒ control for gender, use the treatment

Simpson (1951), in an example similar to this one:
"..., yet it is the combined table which provides what we would call the sensible answer..."

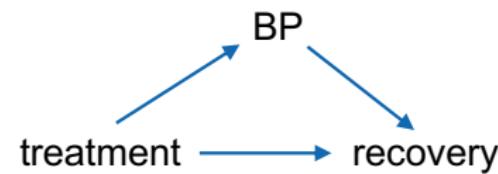
⇒ don't control for BP, don't use the treatment

Simpson's paradox and causal diagrams

- Same numbers, different conclusions...
 - Must use additional information: “story behind the data”, **causal assumptions**
- Consider total causal effect of treatment on recovery
 - Possible scenarios:

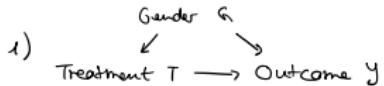


gender is a **confounder**;
control for gender



BP is an **intermediate variable**;
don't control for BP

Notes week 3 - IV



* Observational distribution:

$$P(G, T, Y) = P(G) P(T|G) P(Y|T, G)$$

* Interventional distribution under $do(T=1)$
 ("force all people to take the treatment")

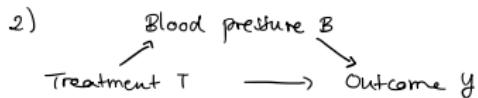
$$P(G, T=1, Y | do(T=1)) = P(G) \mathbb{1}\{T=1\} P(Y | T=1, G)$$

Marginalize:

$$P(G, Y | do(T=1)) = P(G) P(Y | G, T=1)$$

$$P(Y | do(T=1)) = \sum_g P(G=g) P(Y | G=g, T=1)$$

↓ ← both parts known from observation conditional on gender



* Observational distribution:

$$P(Y, T, B) = P(T) P(B|T) P(Y|B, T)$$

* Jnt. distribution:

$$P(y, \tau, B | do(\tau=1)) = \mathbf{1}_{\{\tau=1\}} P(B|\tau) P(y|B, \tau)$$

Marginalize:

$$P(y | do(\tau=1)) = \sum_b P(B=b | \tau=1) P(y | B=b, \tau=1)$$

$$= P(y | \tau=1)$$

also known from unconditional wrt B
observation

Causal DAGs

- Causal DAGs imply strong assumptions, allowing us to estimate post-intervention distributions from observational data
- How do we know the causal DAG?
 - Now: assume it is given, e.g. from background knowledge
 - Later: consider learning causal DAG (under some assumptions)
 - In any case, causal DAG provides clear framework to state causal assumptions for analysis
 - Allows for an honest debate about such assumptions
 - Can draw several possible causal DAGs, conduct the analysis for each of them and perform a sensitivity analysis

Discussion

Any comments or questions?

We may not always find an answer, and since we're not very familiar with causality, we will need to dedicate more time to this topic.