Causality: Models, Reasoning, and Inference

Jonathan Y. Zhou

Georgia Institute of Technology [Yao Group, ISyE]

GT TheoryClub [SP 2024]

The Roosterand the Sun

> The rooster's crow does not cause the sun to rise

Even this simple fact cannot easily be translated into a a statement in formal language/a mathematical equation!

- ...but in fact we shall see how to express the language of *do-calculus*

Overview

Trace the development of causality from a nebulous concept to a precise mathematical theory...

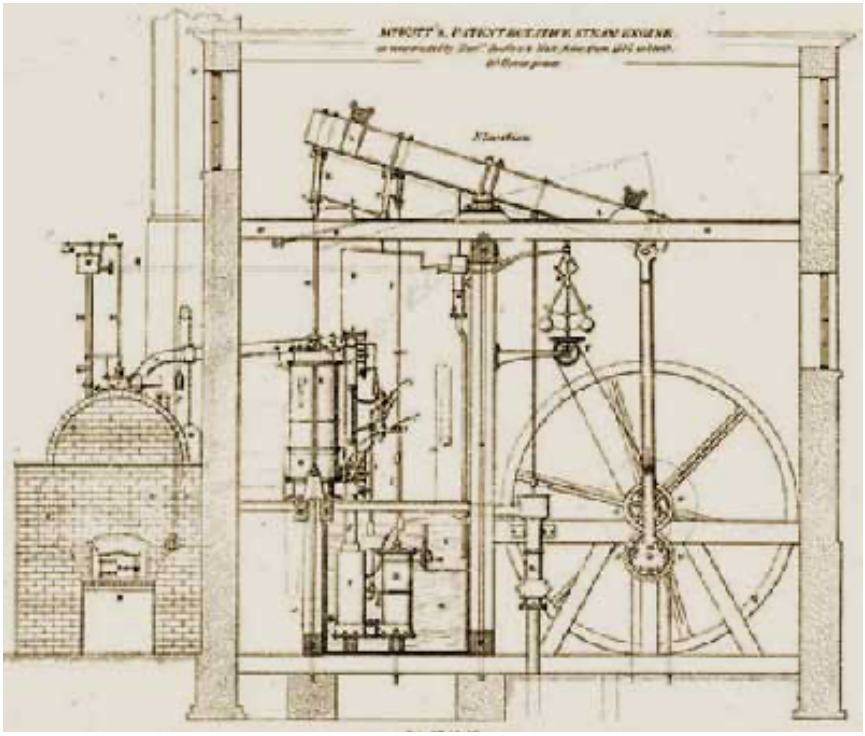
- Historical and philosophical motivation
- Central ideas of causality
- Some tools of causal inference
    - Applications

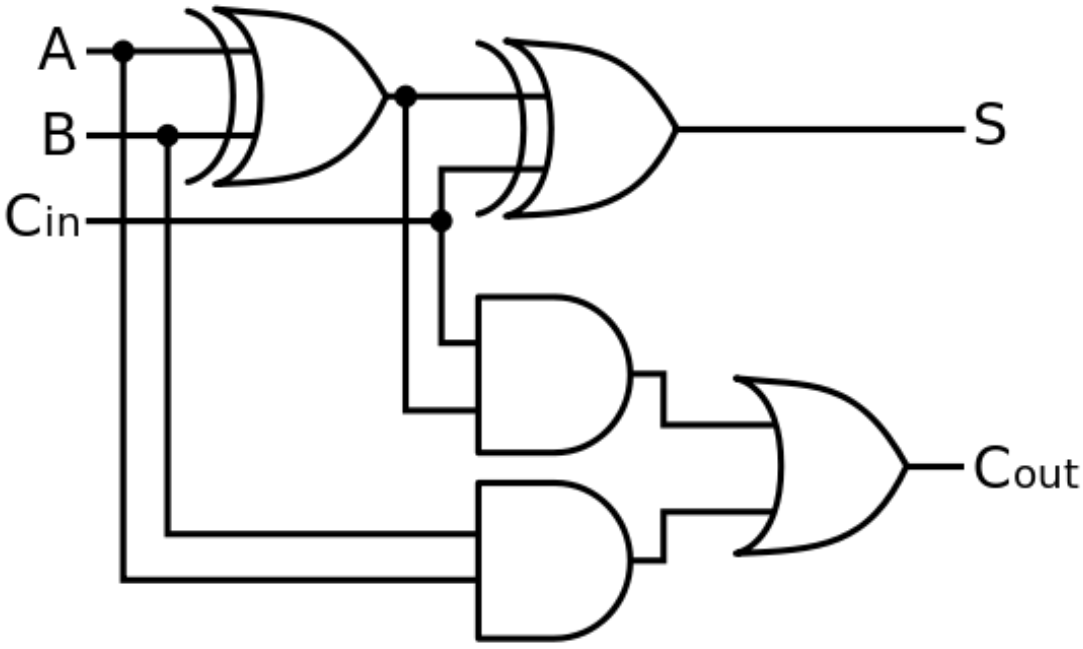
In principio...

Causalityas a HumanConstruction

(L) Juno asks Aeolus to release the winds
(C) Adam and Eve eat from the tree of knowledge
(R) A Roman augur, holding a lituus, next to a sacred chicken

The Renaissance and Industrial Revolution





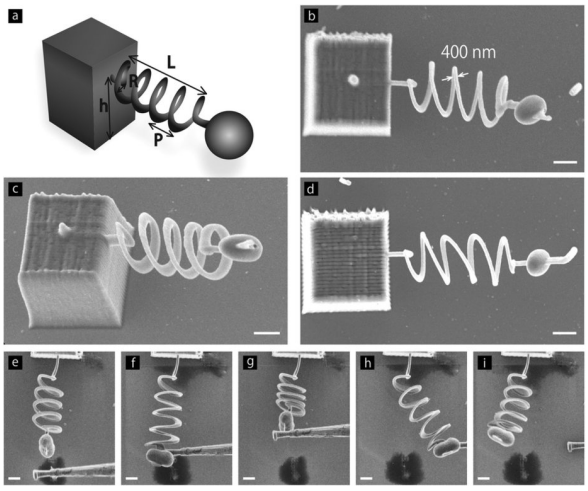
(L) Jame's Watt patent on the steam engine

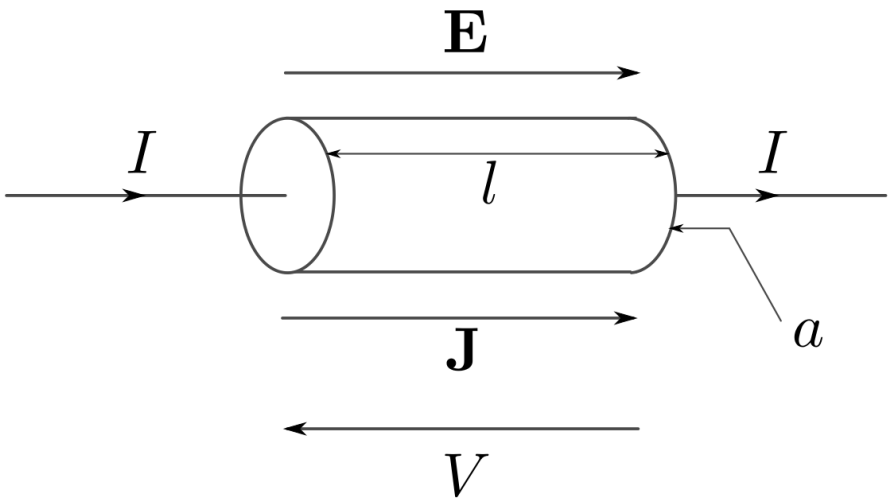
(R) Digital Logic Circuit

Early Science and the Development of Empirical Relations

- Models seek to describe world first, explain second (Galileo)
- Models <-> “laws”
  - Ideally, subject the "laws" to falsification
  - “The mission is to classify truths, not to certify them” (Miller)
  - Since classical antiquity black swans were was presumed not to exist
    - “*rara avis in terris nigroque simillima cygno*” (Juvenalis)
    - Yet black swans were encountered in Australia by Europeans in the 17th century
    - How to deal with of inductive knowledge: connection to modern machine learning (out of domain learning/generalization)

Empirical “Laws”





In[ ]:=
```
FormulaData["OhmsLaw"]
```

Out[ ]=

$$V == I\, R$$

In[ ]:=
```
FormulaData[{"HookesLaw", "PotentialEnergy"}]
```

Out[ ]=

$$U == \frac{1}{2}\, k\, x^2$$

The Characterization of Physical Law and the Riddle of Causation

Physical Law given in the form of PDEs. e.g. Maxwell's equations

$$\nabla \cdot \mathbf{E} = \frac{\rho}{\epsilon_0} \tag{1}$$

$$\nabla \cdot \mathbf{B} = 0 \tag{2}$$

$$\nabla \times \mathbf{E} = \frac{-1}{c} \frac{\partial \mathbf{B}}{\partial t} \tag{3}$$

$$\nabla \times \mathbf{B} = \frac{1}{c} \left(4 \pi \mathbf{J} + \frac{\partial \mathbf{E}}{\partial t} \right) \tag{4}$$

Or even in the "stochastic" world:

$$i\hbar \frac{\partial}{\partial t} \mid \Psi[t] \rangle = \hat{H} \mid \Psi[t] \rangle \tag{5}$$

In the most fundemental physical law, there is NO "CAUSE" and NO "EFFECT".
Classical approximation to the universe states that the universe simply proceeds according predefined trajectory in "state space" (i.e. no free will) ("up-to stochastics")

What is Causation?

The Riddles:

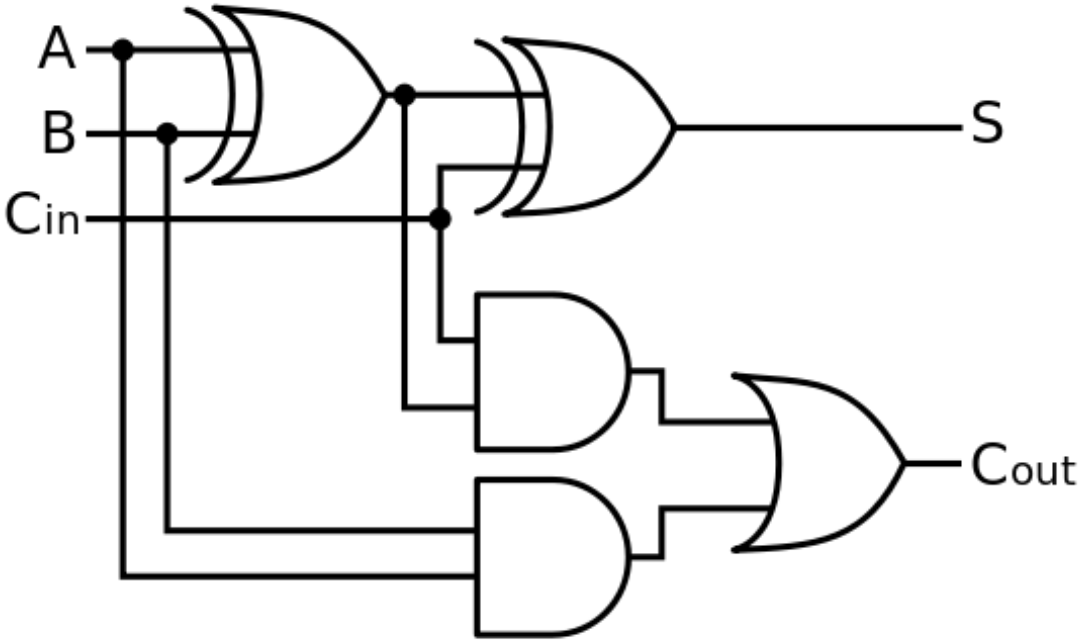
How do people EVER acquire knowledge of CAUSATION?

- "Regularity of succession" necessary, but not sufficient

What DIFFERENCE does it make if I told you that a certain connection is or is not causal?

- *"It could not possibly be an abbreviation, because the laws of physics are all symmetrical, going both ways, while causal relations are uni-directional, going from cause to effect."* (Russell)

Computingand Representation



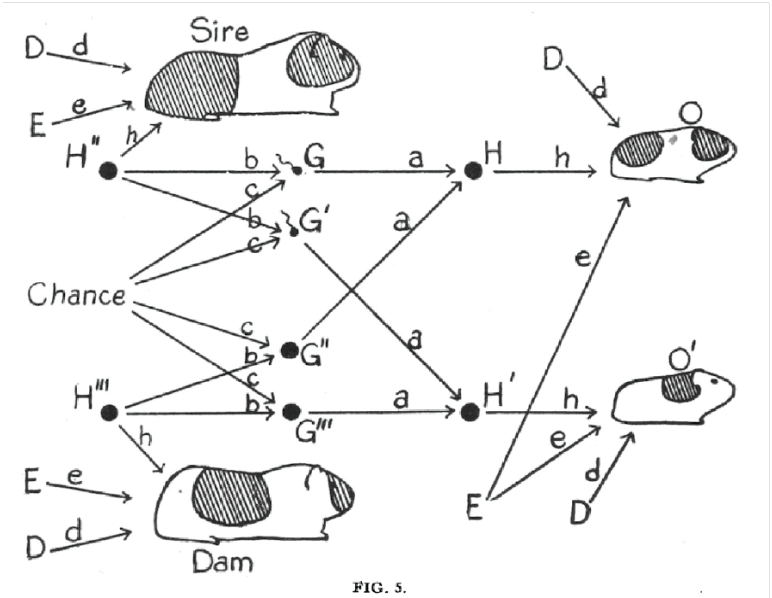
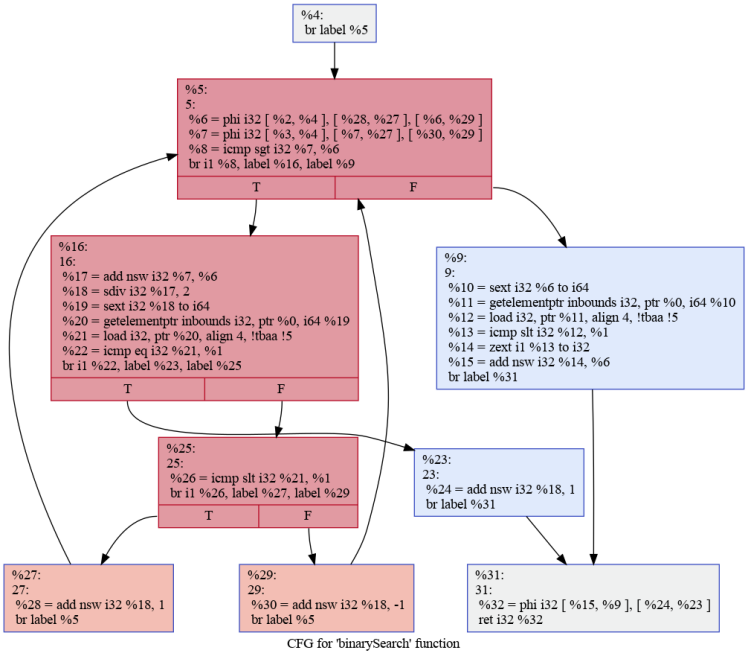


FIG. 5.

Diagram illustrating the casual relations between litter mates (O, O′) and between each of them and their parents. H, H′, H″, H,‴ represent the genetic constitutions of the four individuals, G, G′, G″, and G‴ that of four germ cells. E represents such environmental factors as are common to litter mates. D represents other factors, largely ontogenetic irregularity. The small letters stand for the various path coefficients.



CFG for 'binarySearch' function

(L) Full adder in digital logic [Deterministic]

(C) The phenotypic traits of guinea pigs [Stochastic]

(R) Control flow graph for a binary search procedure

Act II: A Ladder of Causation

Rain and Wetness: Setting

Consider the follow experiment:

Every day, I stand outdoors for one hour. I record the following pair of Bernoulli random variates:

(R, W) = (RainQ WetQ) (6)

$$R = \begin{cases} 0 & \text{No Rain} \\ 1 & \text{Rain} \end{cases} \tag{7}$$

$$W = \begin{cases} 0 & \text{Not Wet} \\ 1 & \text{Got Wet} \end{cases} \tag{8}$$

For simplicity, consider an infinite sample (population) setting

i.e. joint probability measure (R,W) completely known

We can verify that R, W are indeed correlated.

Rain and Wetness: Correlation

$$(R, W) = (\text{RainQ}\ \text{WetQ}) \tag{9}$$

Q: R, W are correlated? Check:

$$\Pr[R = 0 \mid W = 0] \stackrel{?}{=} \Pr[R = 0 \mid W = 1] \tag{10}$$

Or (symmetrically)

$$\Pr[W = 0 \mid R = 0] \stackrel{?}{=} \Pr[W = 0 \mid R = 1] \tag{11}$$

In fact these are *testable* from infinite sample setting (i.e. computable in closed form). Or from *sufficient* amount of finite data with high probability, (i.e. as classical problem of *hypothesis testing*)

Rain and Wetness: Causality

Now consider the statement: **Q:** *"Wetness **causes** rain"*

Can we test from observational data? (finite/infinite sample case)

Rain and Wetness: Causality

Now consider the statement: ***Q:*** *"Wetness* ***causes*** *rain"*

Can we test from observational data? (finite/infinite sample case)

Answer is seems to be negative

But if we could do run experiments (*interventions*) to verify the following...

1. Wetness causes rain
2. Rain causes wetness

Rain and Wetness: Causality and Experiment

(A) “Wetness causes rain”
(B) “Rain causes wetness”

The daily experiments:

1. I stand outside under an umbrella (regardless of the weather condition)
2. I stand outside and do “bucket challenge” (regardless of the weather condition)
3. I stand outside and my friends use cloud seeder to force it to always rain
4. I use a giant fan to blow away all the clouds (so the weather is always clear)

Rain and Wetness: Causality and Experiment

Wetness -> Rain

1. I stand outside under an umbrella (regardless of the weather condition)

2. I stand outside and do “bucket challenge” (regardless of the weather condition)

So now check

$$Pr^{(1)}[R = 1 \mid W = 0] \overset{?}{=} Pr^{(2)}[R = 1 \mid W = 1] \tag{12}$$

Notice that we are using different probability measures $Pr^{(1)}$, $Pr^{(2)}$***!***

If the two quantities are equal, then we know that wetness doesn’t cause rain.

Rain and Wetness: Causality and Experiment

Rain -> Wetness

1. I stand outside and my friends use cloud seeder to force it to always rain

2. I use a giant fan to blow away all the clouds (so the weather is always clear)

So now check

$$Pr^{(1)}[W = 1 \mid R = 1] \stackrel{?}{=} Pr^{(2)}[W = 1 \mid R = 0] \tag{13}$$

In this case, we would more than likely see that the two probabilities are different... i.e. that it seems that rain has a causal effect on wetness

Key Ideas:

1. *Correlations* are testable from *observational* data
2. Causality defined and tested through *interventional* data
   - i.e. by actively modifying a value to some prescribed value

The key tools

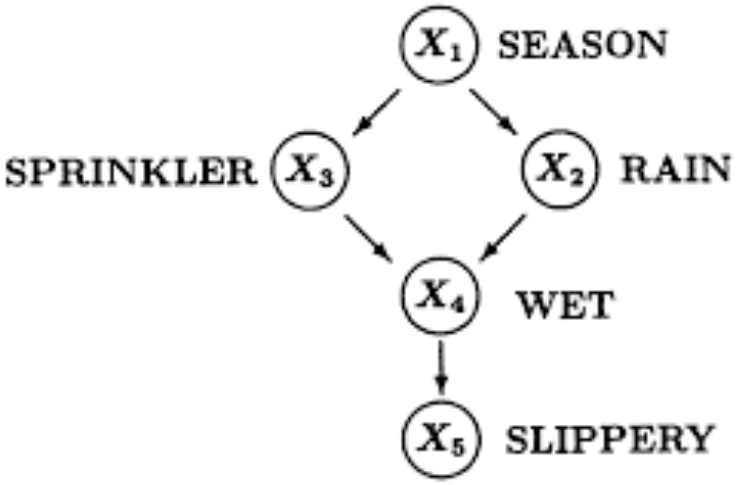
What we will develop is the following [which solves the second riddle, what difference it makes to have causal knowledge]

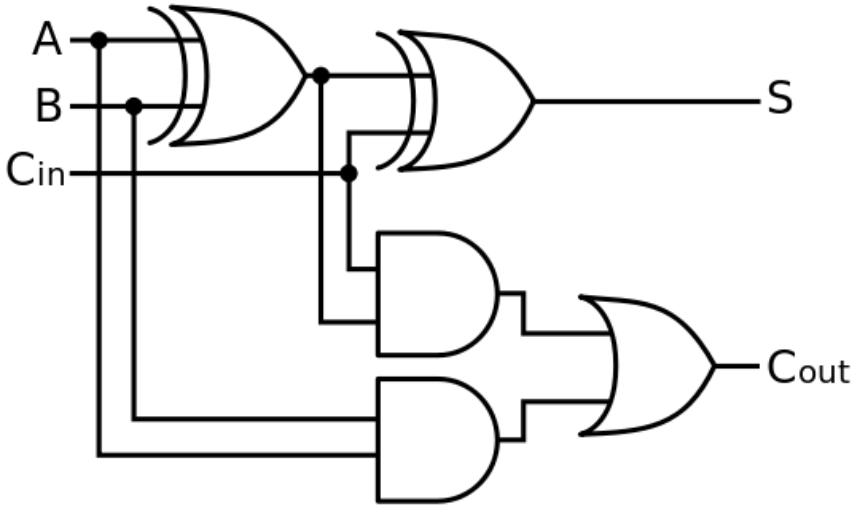
1. A method of representing a (stochastic) system [Directed Graphical Model]
   1.1. System consists of (a) set of variates (b) relations between the variates
2. A method to “jump between” the the class probabiity measures representing interventions on a system [do-calculus]

ProbabilisticGraphicalModels:BayesNet

Consider a joint distribution between two variables. Bayesian network gives an efficient factorization based on "independency assertions"

Causal-DAG encodes more: (1) joint distribution (2) behavior under intervention





$$\Pr[X_1, X_2, X_3, X_4, X_5] =$$

$$\Pr[X_1]\Pr[X_3|X_1]\Pr[X_2|X_1]\Pr[X_4|X_3,X_2]\Pr[X_5|X_4]$$

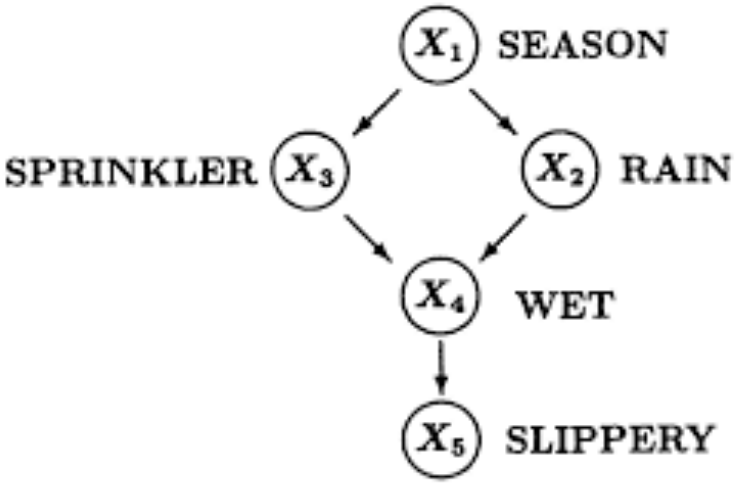
Independence Assertions and Bayesian Networks

Before we discuss further the interventional compatibility, the let's turn to the probabilistic background.

All observations of a system give a *joint distribution* over system variables

Qn: For $X_1 \ldots X_n$ discrete (continuous) r.v.s, how many numbers (dimensions) do I need to describe the joint distribution?

.... but oftentimes, we have more structure. Encoded by DAG.



$$P[X_1, X_2, \ldots, X_n] = \prod_{i=1}^{n} P[X_i \mid Pa[X_i]]; \tag{15}$$

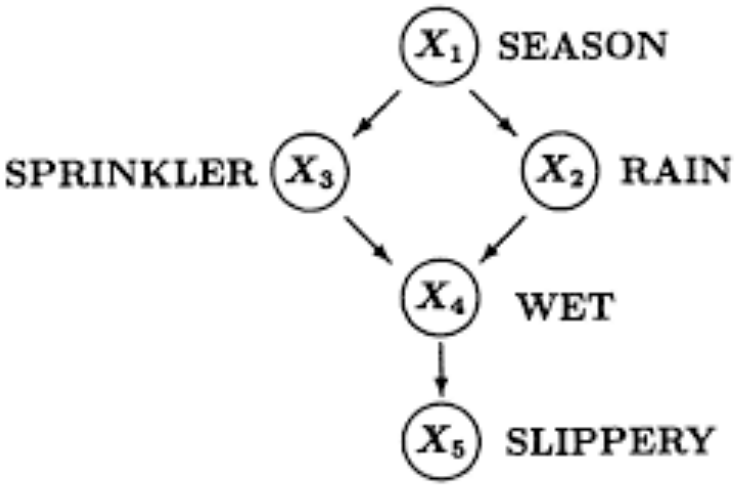
$$\Pr[X_1,X_2,X_3,X_4,X_5] = \Pr[X_1]\Pr[X_3|X_1]\Pr[X_2|X_1]\Pr[X_4|X_3,X_2]\Pr[X_5|X_4] \tag{16}$$

StructuralCausalModel:Generativeviewon BN

Consider a topological sorting over nodes: Then we have equations:

$$X_i := f(\mathrm{Pa}(X_i)) \tag{17}$$

Which gives a sampling procedure for observations of our system. (System Dynamics) **Example:**



$$\begin{aligned} x_1 &\sim P[X_1] \\ x_2 &\sim \mathrm{P}[X_2 \mid x_1] \\ x_3 &\sim \mathrm{P}[X_3 \mid x_1] \\ x_4 &\sim \mathrm{P}[X_4 \mid x_3, x_2] \\ x_5 &\sim \mathrm{P}[X_5 \mid x_4] \end{aligned} \tag{18}$$

Economists call this type of representation: **Structural Equation Modeling**

Remark: You need to assume that the underlying BN is true to the system... Describing the system is often task of domain expert... i.e., econometrist, engineer, biologist who knows/asserts the independencies of the problem. Also some more concerns... faithfulness, and multiple DAGs can be compatible with a given probability distribution.

D-Separationand MarkovConditions:

Recall that for sets A_i, A_j, A_k

$$A_i \perp A_j \mid A_k \leftrightarrow \Pr[A_i \cap A_j \mid A_k] = \Pr[A_i \mid A_k]\Pr[A_j \mid A_k] = \phi_{ik}(x_i, x_k)\ \phi_{jk}(x_j, x_k) \tag{19}$$

DAG allows us to "read off" conditional independence relations of random variables.

Theorem [Global Markov Property]: *i.e. if DAG* G *is compatible with probability measure* Pr, then d-separation corresponds to

$$A_i \perp_G A_j \mid A_k \to A_i \perp A_j \mid A_k \tag{20}$$

Where $A_i \perp_G A_j \mid A_k$ represents “d-separation” of A_i , A_j via A_k.

D-Separation:Definition

Given a DAG G, and sets of random variables A_i , A_j , A_k.

Defn: A_k d-separates A_i , A_j under G [$A_i \perp_G A_j \mid A_k$] if

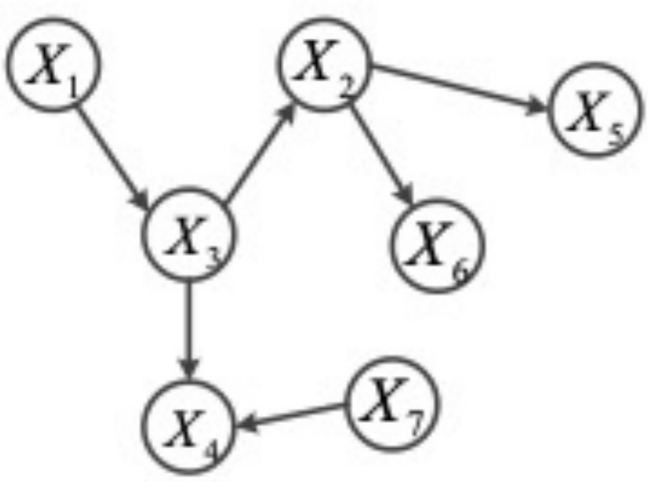
All paths between $a \in A_i$, $b \in A_j$ are blocked by A_k.

Defn: A path between $a \in A_i$, $b \in A_j$ is blocked by A_k if

Some length 3 path (x-y-z) along a...->...b is blocked by A_k, i.e.

- [x->y->z] y $\in A_k$ (Causal Trail)
- [x<-y<-z] y $\in A_k$ (Evidential Trail)
- [x<-y->z] y$\in A_k$ (Common Cause)
- [x->y<-z] y $\notin A_k$ (Common Effect, Collider, Immorality)

D-Separation: Example



- $X_1 \perp_G X_5 \mid X_2$?
- $X_2 \perp_G X_7 \mid X_4$?
- $X_4 \perp_G X_5 \mid X_3$?

Interventions as Surgery on SCMs

Observational SCM

Let $(X_1, X_2, \ldots X_d)$ be a collection of random variables defined by SCM C. Let j be an index over the r.v.s, let Pa denote the parents at index j, and N a noise random variables at index j, such that the noise terms are mutually independent.

$$x_j := f_j(\text{Pa}[j], \text{N}[j]) \; ; \; j \in [d] \tag{21}$$

Let the probability measure induced on $(X_1, X_2, \ldots X_d)$ be $\Pr^C[\cdot]$

Intervention SCM and Intervention Distribution

Let $\tilde{C}$ be the SCM associated with replacing one or more structural equations associated with C

i.e. replace some $x_j := f_j(\text{Pa}[j], N[j])$ with

$$x_j := \tilde{f}_j(\tilde{\text{Pa}}[j], \tilde{N}[j]) \tag{22}$$

i.e. with a new function, parents, and noise terms. The $\tilde{C}$ induces a new probability measure $\Pr^{\tilde{C}}[\cdot]$, which is related (but distinct from) the original SCM

$$\Pr^{\tilde{C}}[\cdot] = \Pr^{C; \text{do}\left[x_j := \tilde{f}_j(\tilde{\text{Pa}}[j], \tilde{N}[j])\right]}[\cdot] \tag{23}$$

Notice that this also induces a new DAG associated with $\tilde{C}$.

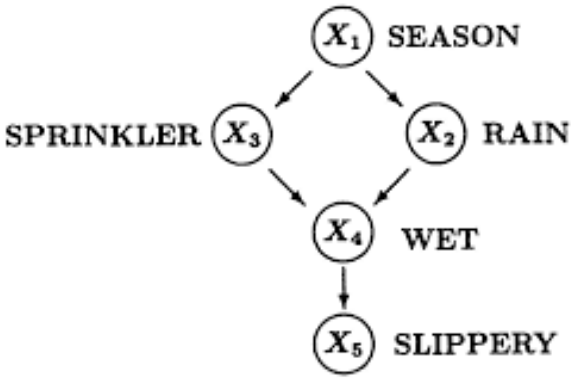
Example: Atomic intervention

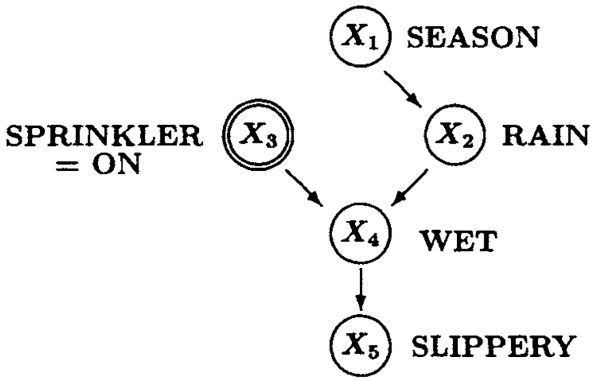
$$x_j := a \tag{24}$$

Interventionas Surgeryon DAGs

Example: Atomic intervention.

$$x_3 := \text{ON} \tag{25}$$





Graphically: Delete all parents going into X_3, then set X_3 to on. Outgoing edges unaltered

Remark: In general ***do is not conditioning***

$$\Pr^{C}[\cdot \mid X_{3} = \text{On}] \neq \Pr^{C;\, \text{do}[X_3 := \text{On}]}[\cdot] \tag{26}$$

Remark: This is the type of intervention we have in *controlled experiments!*

CausalEffectandTreatmentEffects

Definition: Total Causal Effect. A total causal effect exists from X_i to X_j if and only if there exists a random variable $\tilde{N}_{x_i}$ such that

$$X_i \not\perp\!\!\!\perp X_j \text{ under } \Pr^{C;\, \mathrm{do}\left(X_i := \tilde{N}_{x_i}\right)} \tag{27}$$

Remark: The above connects causality with *independence on the interventional measure.* Furthermore note that the relation is directional.

Remark: If there does not exist a directed path from X_i to X_j in the DAG associated with C, then there is no TCE from X_i to X_j

The Calculus for Interventions

Turing Award (Pearl, 2011)

Setting: Known SCM C = (S, N_x) and an associated Markov DAG G=(V,E) which is compatible with joint distribution ($\boldsymbol{X_1}$,$\boldsymbol{X_2}$,...$\boldsymbol{X_d}$)

Suppose that an intervention $X_j := \tilde{N}_j$ occurs.

Goal: Compute interventional distribution but use only the observed distribution. i.e. compute $\mathrm{Pr}_Y{}^{C;\,\mathrm{do}(X:=x)}(y)$ from only observational distribution. (First Question: Identifiably of an interventional distribution, does there exist a way?)

Tool: Do-Calculus: Provides a set of rules that allows for the manipulation of a conditional distribution in the interventional SCM. [Tran 2002, gives an algorithm]

Casual Inference in the Observational Setting

In many cases (i.e. drug testing, policy research), it is not feasible (or highly unethical) to conduct a double-blind experiment.

Is there any hope to learn any causal knowledge?

Casual Inference in the Observational Setting

In many cases (i.e. drug testing, policy research), it is not feasible (or highly unethical) to conduct a double-blind experiment.

Is there any hope to learn any causal knowledge?

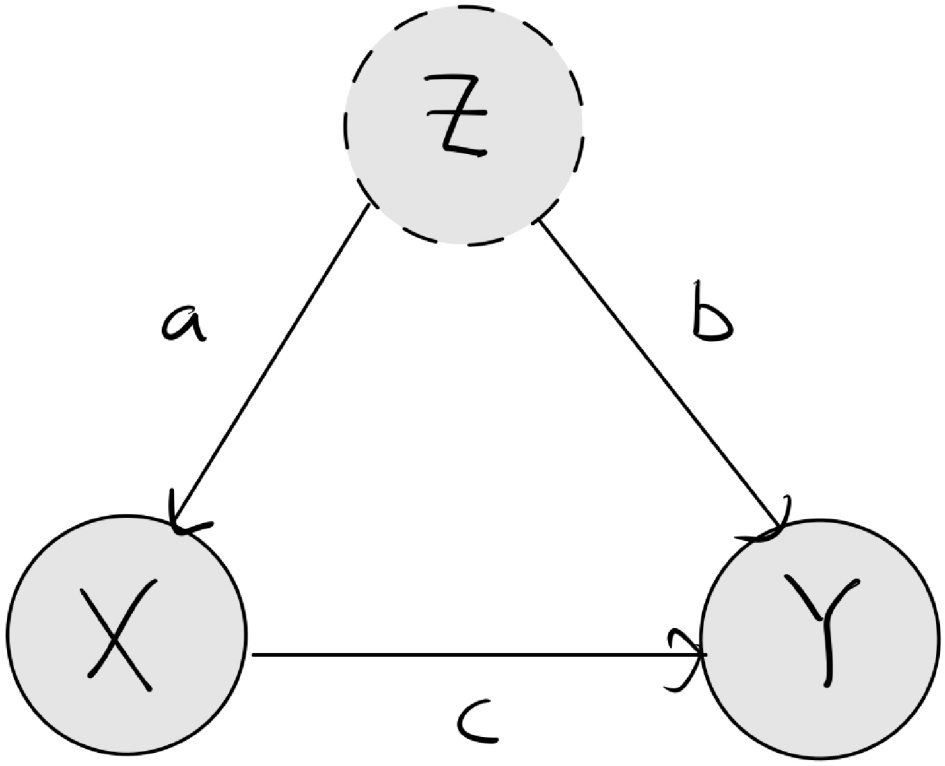
At first glance... observation gives only correlations...
But in fact, **yes**... if (1) the situation is right ("Natural Experiments") and (2) you are "clever enough"

I will present one method... method of instrumental variables

- Nobel Prize in Economic Sciences: (Card, Angrist, and Imbens, 2021)

The Setting: Regression with Confounders

Consider the following causal model with known parameteric form



$$\begin{aligned} Z &:= N_1 \\ X &:= aZ + N_2 \\ Y &:= cX + bZ + N_3 \\ N_1, N_2, N_3 &\sim N(0, 1)\ \text{i.i.d} \end{aligned} \tag{28}$$

Goal is to recover the relation X->Y (i.e. value of c) from data.
Point: You do not get to see Z, so it is not possible to determine c.

Analysis:

Consider the following program regressing X on Y to recover c

$$\min_\gamma \mathrm{E}\left[(Y - \gamma X)^2\right] \tag{29}$$

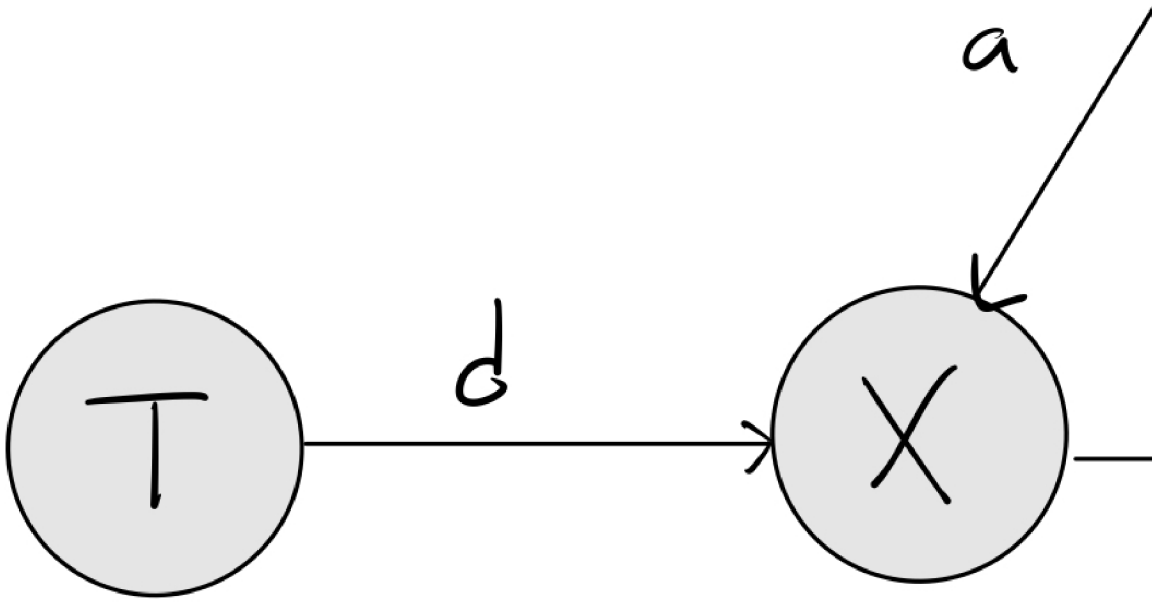
The closed form solution (immediate from KKT condition) is

$$\gamma^* = \frac{E[XY]}{E[X^2]} = \frac{c\,E[X^2] + b\,E[XZ]}{a^2\,E[Z^2] + E[N_2{}^2]} = c + b\left(\frac{a^2}{a^2+1}\right) \tag{30}$$

Which is a biased estimate of c. So with nonzero latents (b, a), we cannot recover c without bias. In this setting

X is called endogenous

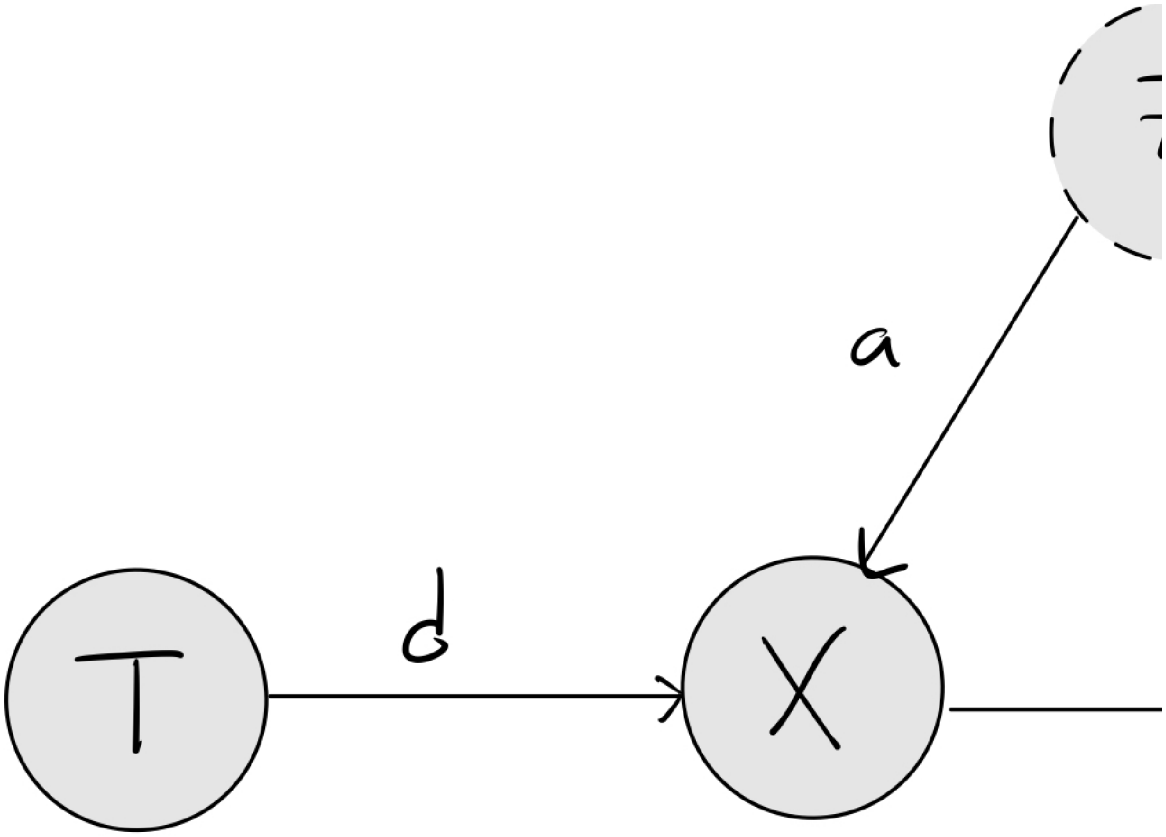
The Method of Instrumental Variables



Suppose we have another variable T (Instrumental Variable):

- Such that T affects X [Relevancy]
- But T does not directly influence any other variable [Unconfoundedness]

Example:



X = Smoking

Y = Lung Cancer

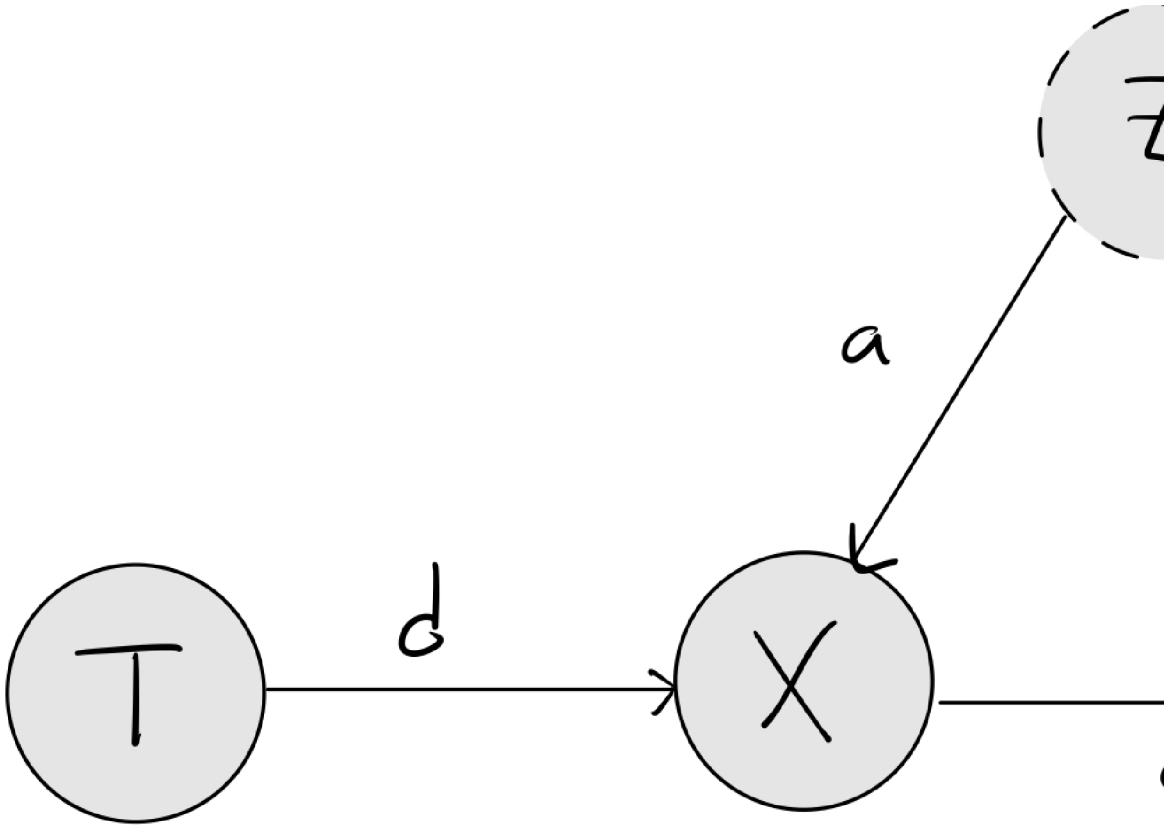
T = Tax on Cigarettes

Z = Confounders (e.g. Depression, Social Influence, Genetics)

Tobacco companies came up with the explanation that the confounding factors were the common cause to smoking and lung cancer, and asserted that there is *no casual link* smoking->cancer.

[R: Pearl, Art and Science of Cause and Effect]

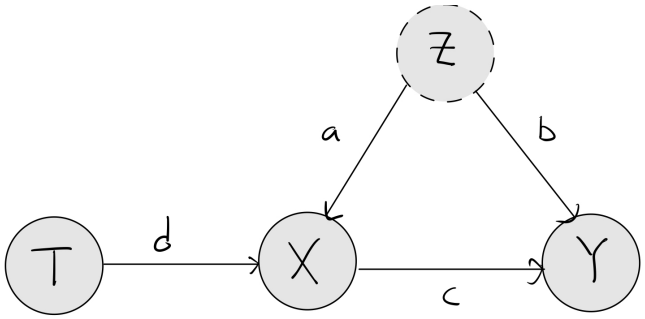
Example:



X = Smoking
Y = Lung Cancer
T = Tax on Cigarettes
Z = Confounders (e.g. Depression, Social Influence, Genetics)

Method of IV gives us a method to understand the link X->Y. Suppose that the tax on cigarettes does not influence the unobserved factors (Z). Then we can estimate as follows (Leigh and Schembri, 2004):

IV 2SLS Estimation:



$$T = N_4 \sim N[0, 1]; \quad X = aZ + dT + N_2 \tag{31}$$

Suppose the effect of T on X is linear (with gaussian noise), notice that

$$d = \alpha^* = \text{argmin}_\alpha \, E\left[(X - \alpha T)^2\right] \tag{32}$$

i.e. we can use (T,X) to give an unbiased estimate of d.

$$\beta^* = \text{argmin}_s \, E\left[(Y - sT)^2\right] \tag{33}$$

We know that

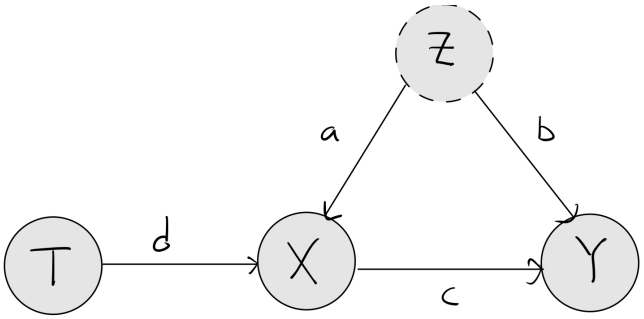
$$\beta^* = \frac{E[TY]}{E\left[T^2\right]} = \frac{cE[TX] + bE[TZ] + E[N_3\, T]}{E\left[T^2\right]} = \frac{cdE\left[T^2\right]}{E\left[T^2\right]} = c\, d \tag{34}$$

So then we know that

$$c^* = \frac{\beta^*}{\alpha^*} \tag{35}$$

is an unbiased estimator of c

AnotherExample



Does military service result in future earnings? (Angrist 1990)

`X = Military Service`

`Y = Future Earnings`

`T = Vietnam Draft Number`

`Z = Confounders`

Relevance: Vietnam Draft Number directly determines military service

Unconfoundedness: T is randomly selected, and so does not influence Z or Y directly.

Outlookand Perspective

Doesthe roostercrow causethe sun to rise? [C: R->S]

$$\mathrm{Pr}^{C;\,\mathrm{do}\,(R:=0)}[S=1]-\mathrm{Pr}^{(C;\,\mathrm{do}\,(R:=1))}[S=1]\stackrel{?}{=}0 \tag{36}$$

What we covered:

- Historical development of causation
- A mathematical framework for causality
- Causality from experiment
- Causality from observation

Perspective

- So far we supposed that the underlying causal DAG is (1) known and (2) has parametric form
  - Even stronger for IV: we assumed that our variables have linear relationship with gaussian noise.
    - Vector case: easy
    - **Fact:** Must assume known parametric form "*No Nonparametric Identification of the ATE*"
    - Nonlinear case, may lose guarantees (but often GLM is ok).
  - IV approach derivable from do-calculus, similar analysis can be done in more general graph settings.
  - Goal: Recovery parameters: **Causal Inference**
- In many situations (e.g. Biology: *pathways*) our task is to recover a DAG from a (very large) list of possible candidate interactions.
  - Experiments are expensive
  - Naively, the number of DAGs as a function of the number of nodes, $G(n)$, is super-exponential in n.
  - But we know some independencies.
  - **Causal Discovery**

ReferencesandFurtherReading

- Causality: Models, Reasoning, and Inference (Pearl)
- The Art and Science of Cause and Effect (Pearl)
- Elements of Causal Inference: Foundations and Learning Algorithms (Peters, Janzing, and Schölkopf)
- Mostly Harmless Econometrics (Angrist and Pischke)
- Short Course on Causal Inference (Shakkottai)
- Probabilistic Machine Learning: Advanced Topics (Murphy)
- Probabilistic Graphical Models (Koller and Friedman)