Causal Inference & Causal Learning

DAG

Xiaolei Lin

School of Data Science
Fudan University

April 5, 2024

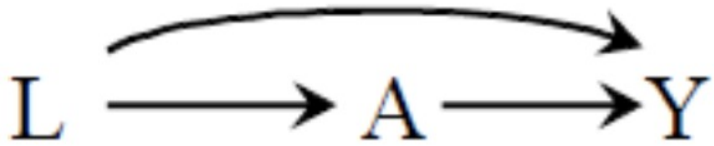
Recap

- g-formula in Randomized experiments and Observational studies
- Effect modification
- Interaction

Today's plan

- Directed Acyclic Graphs
- Collider and mediator
- Block and d-separation

Directed Acyclic Graphs (DAGs)

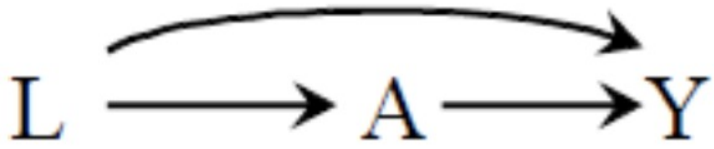


- DAG: A directed acyclic graph (or DAG) is a set of nodes and edges amongst nodes, each of which is directed (only goes in one direction) such that there is no directed path from a node back to itself

Directed Acyclic Graphs (DAGs)

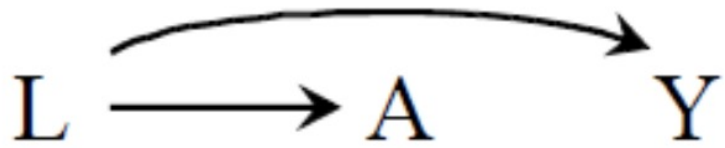
- Associated with each graph is an underlying causal model
- Useful to classify sources of bias and to identify potential problems in study design and analysis
- Some find it to be more intuitive than the counterfactual approach, however two approaches are intimately linked
- Modern theory of causal diagrams mainly arose within the disciplines of computer science and artificial intelligence

Directed Acyclic Graphs (DAGs)



- The diagram comprises three nodes representing random variables (L, A, Y) and three edges (the arrows)
- We shall adopt the convention that time flows from left to right, and thus L is temporally prior to A and Y , and A is temporally prior to Y
- This is a Directed Acyclic Graph (DAG)
  1. It is directed because all edges are arrows
  2. It is acyclic because there are no directed cycles

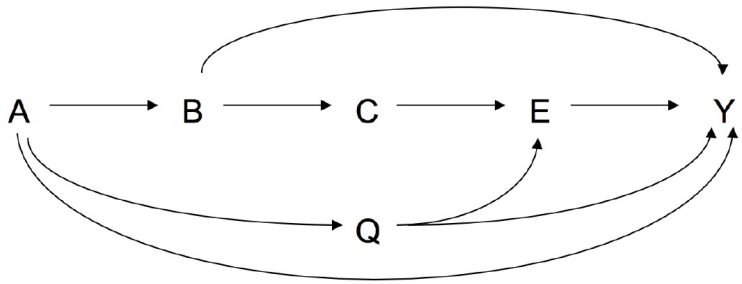
Directed Acyclic Graphs (DAGs)



- Presence of an arrow from, say, A to Y indicates either a direct causal effect (i.e. not mediated by any variable on the graph) for at least one person, or that we cannot assume such individual causal effects do not exist

- Lack of an arrow would mean that we know or can assume that there is no direct causal effect of A on Y for any individual in the population

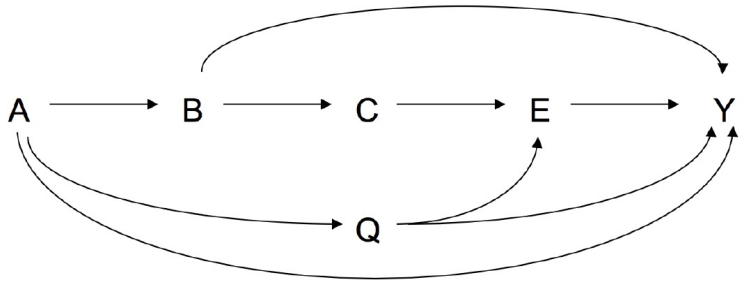
Directed Acyclic Graphs (DAGs)



For a DAG to be CAUSAL we require:

1. The arrows of a causal DAG represent causal or counterfactual relations. E.g. The Q-E arrow is present if after intervening on C (the other node which is a parent of E) it is still possible to affect E by intervening on Q.
2. All "common causes" of any two variables on the DAG must also be on the DAG

Directed Acyclic Graphs (DAGs)



- Path: Any sequence of nodes connected by edges
- Directed Path: A path which is such that the order of its nodes follows the direction of the edges
- Collider: A node on a path with both arrows on the path going into that node (e.g. E on the path $A \rightarrow Q \rightarrow E \leftarrow C$)
- Note: A collider is relative to the path: E is a collider on $A \rightarrow Q \rightarrow E \leftarrow C$, but not on $A \rightarrow Q \rightarrow E \rightarrow Y$

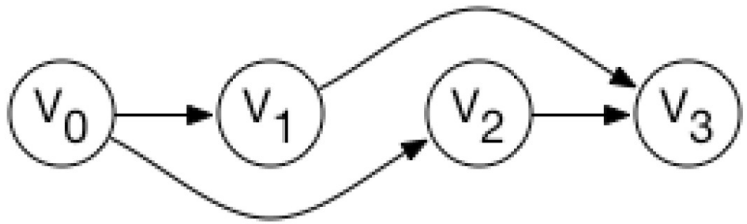
Directed Acyclic Graphs (DAGs)

- Standard causal diagram does not distinguish whether an arrow encodes a harmful effect or a protective effect
- Standard causal diagram does not encode how two causes interact
- DAGs have applications other than causality. Here we only consider causal DAGs
- We shall say that a graph is causal if common causes of any pair of variables in the graph are also in the graph

Directed Acyclic Graphs (DAGs)

- While causal effect follows direction of arrows along a path, associations do not and are symmetric
- A path between two variables in a DAG is a route that connects the two variables by following a sequence of (nonintersecting) edges
- A path is causal if it consists entirely of edges with their arrows pointing in the same direction, otherwise it is noncausal

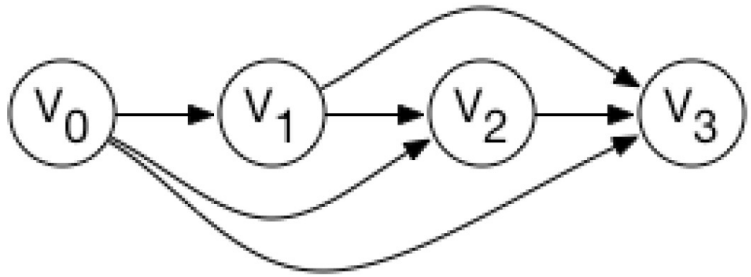
More notations on DAG



We use PA_j to denote the parents of V_j, i.e., the set of nodes from which there is a direct arrow into V_j.

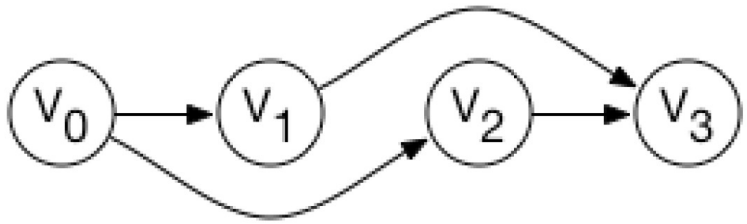
- $PA_0 = \emptyset$, $PA_1 = V_0$, $PA_2 = V_0$, $PA_3 = \{V1, V_2\}$

Complete DAG



- A DAG is called complete if there is an arrow between every pair of nodes.
- Given an ordered collection of random variables $V = \bar{V} = (V_0, \ldots, V_m)$, a DAG is complete if:
  - For $i < j$, there is an arrow $V_i \to V_j$.
  - Equivalently, $PA_i = \{V_0, \ldots, V_{i-1}\}$.

Statistical DAG



We say that a DAG represents the joint density of its node variables V if, and only if, it satisfies the Markov factorization:

$$f(\nu) = \prod_{j=1}^{M} f(\nu_j \mid pa_j)$$

In this case, we have

$$f(\nu) = f(\nu_3 \mid \nu_1, \nu_2) f(\nu_2 \mid \nu_0) f(\nu_1 \mid \nu_0) f(\nu_0)$$

Statistical DAG

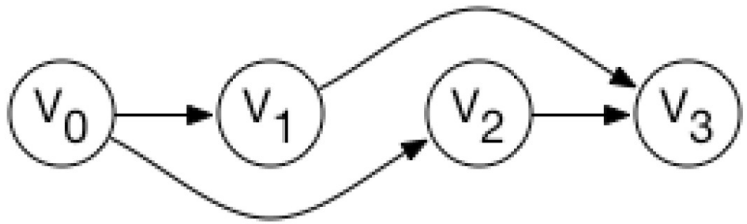
We say that a DAG represents the joint density of its node variables V if, and only if, it satisfies the Markov factorization:

$$f(\nu) = \prod_{j=1}^{M} f(\nu_j \mid pa_j)$$

This is equivalent to the statements:

- Each variable is independent of all variables in its past, given its parents
- Each variable is independent of its nondescendants, given its parents

Statistical DAG

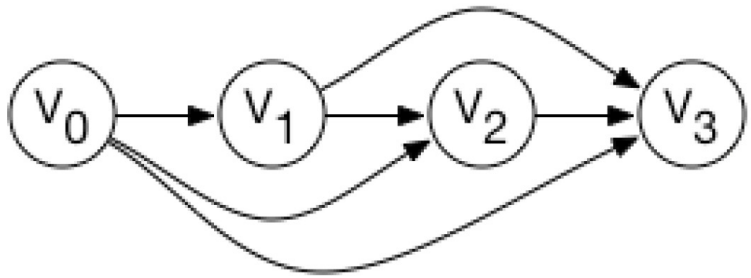


Incomplete DAG

$$f(\nu) = f(\nu_3 \mid \nu_1, \nu_2) f(\nu_2 \mid \nu_0) f(\nu_1 \mid \nu_0) f(\nu_0)$$

- V_2 is independent of V_1 given (within each level of) V_0.
- or, given V_0, the probability that V_2 takes any value is not predicted by V_1.

Statistical DAG



Complete DAG

$$f(\nu) = f(\nu_3 \mid \nu_0, \nu_1, \nu_2) f(\nu_2 \mid \nu_1, \nu_0) f(\nu_1 \mid \nu_0) f(\nu_0)$$

DAGs and Models

- The statistical model associated with a DAG is the set of distributions it represents.
- The complete DAG with no missing arrows is a nonparametric (saturated) model.
- The complete DAG puts no restriction on the probability distributions it represents.
- Incomplete DAGs can only represent probability distributions that are compatible with their Markov Factorization.

DAGs and Marginal independence

Aspirin use A has a preventive causal effect on the risk of heart disease Y

Represent in potential outcome framework:

$$Pr(Y^{a=1} = 1) \neq Pr(Y^{a=0} = 1)$$

Represent using DAG

DAGs and Marginal independence

Aspirin use A has a preventive causal effect on the risk of heart disease Y

- If A has a causal effect on Y , then generally A and Y are associated
- With unconditional exchangeability, causation $Pr[Y^{a=1} = 1] \neq Pr[Y^{a=0} = 1]$ implies association $Pr[Y = 1 \mid A = 1] \neq Pr[Y = 1 \mid A = 0]$, and vice versa
- Association, unlike causation, is a symmetric relationship between two variables
- association flows between two variables regardless of the direction of the causal arrows

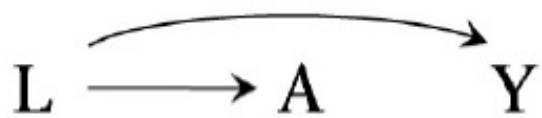
DAGs and Marginal independence

Carrying a lighter A has no causal effect (causative or preventive) on anyone's risk of lung cancer Y, while cigarette smoking L has a causal effect on both carrying a lighter A and lung cancer Y

Represent in potential outcome framework:

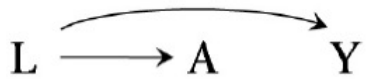
$$Pr(Y^{a=1} = 1) = Pr(Y^{a=0} = 1)$$
$$Pr(Y^{l=1} = 1) \neq Pr(Y^{l=0} = 1)$$
$$Pr(A^{l=1} = 1) \neq Pr(A^{l=0} = 1)$$

Represent using DAG



DAGs and Marginal independence

Is carrying a lighter A associated with lung cancer Y?



- we know $Pr(Y^{a=1} = 1) = Pr(Y^{a=0} = 1)$, is it also true that $Pr(Y = 1 \mid A = 1) = Pr(Y = 1 \mid A = 0)$?
- if a person is carrying a lighter $(A = 1)$, then it is more likely that he/she is a smoker $(L = 1)$, and therefore he/she has a greater than average risk of developing lung cancer $(Y = 1)$.
- having information about the treatment A improves our ability to predict the outcome Y , even though A does not have a causal effect on Y
- L: common cause, confounder

DAGs and Marginal independence

Genetic haplotype A has no causal effect on anyone's risk of becoming a cigarette smoker Y, while both the haplotype A and cigarette smoking Y have a causal effect on the risk of heart disease L.

Represent in potential outcome framework:

$$Pr(Y^{a=1} = 1) = Pr(Y^{a=0} = 1)$$
$$Pr(L^{a=1} = 1) \neq Pr(L^{a=0} = 1)$$
$$Pr(L^{y=1} = 1) \neq Pr(L^{y=0} = 1)$$

Represent using DAG

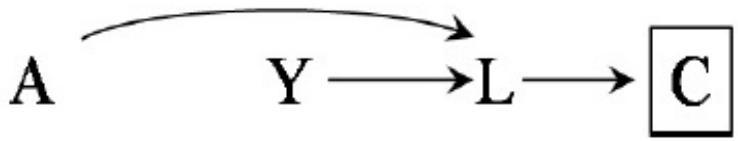
DAGs and Marginal independence

Is genetic haplotype A associated with cigarette smoking Y?

- we know $Pr(Y^{a=1} = 1) = Pr(Y^{a=0} = 1)$, is it true that $Pr(Y = 1 \mid A = 1) = Pr(Y = 1 \mid A = 0)$ or $Pr(Y = 1 \mid A = 1, L) = Pr(Y = 1 \mid A = 0, L)$?
- If a person does not have the haplotype $(A = 0)$, Is he/she more or less likely to become a cigarette smoker $(Y = 1)$ than the average person?
- If a person who has heart disease $L = 1$ does not have the haplotype $(A = 0)$, Is he/she more or less likely to become a cigarette smoker $(Y = 1)$ than the average person?
- L: common effect, collider

DAGs and Marginal independence

Genetic haplotype A has no causal effect on anyone's risk of becoming a cigarette smoker Y, while both the haplotype A and cigarette smoking Y have a causal effect on the risk of heart disease L, C is a consequence of L.



Is genetic haplotype A associated with cigarette smoking Y conditional on C?

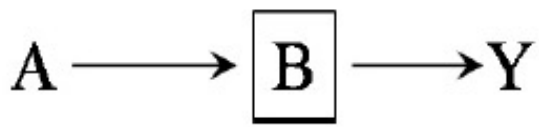
DAGs and Marginal independence

Aspirin A affects the risk of heart disease Y because it reduces platelet aggregation B

Represent in potential outcome framework:

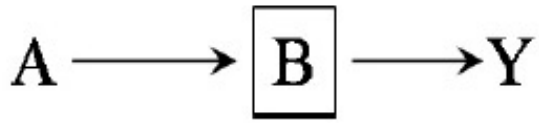
$$Pr(Y^{a=1} = 1) \neq Pr(Y^{a=0} = 1)$$
$$Pr(B^{a=1} = 1) \neq Pr(B^{a=0} = 1)$$
$$Pr(Y^{b=1} = 1) \neq Pr(Y^{b=0} = 1)$$

Represent using DAG



DAGs and Marginal independence

Is there an association between A and Y within levels of (conditional on) B?



- for individuals with low platelet aggregation $B = 0$, regardless of whether he/she is treated ($A = 1$ or $A = 0$), he/she has a lower than average risk of heart disease
- aspirin use affects heart disease only through platelet aggregation, learning an individual's treatment status does not contribute any additional information to predict risk of heart disease
- B: mediator; A and Y are marginally associated, but conditionally independent given B

DAGs and association

Summary: on a causal DAG, statistical association between two variable X and Y can arise in one of three ways:

1. X causes Y (or Y causes X)
2. X and Y have a common cause C
3. X and Y have a common effect K and we are calculating our measure of association conditional on K. K is also called "collider"

Block

A path is either blocked or open according to the following graphical rules

1. If there are no variables being conditioned on, a path is blocked if and only if two arrowheads on the path collide at some variable on the path. e.g., the path $L \rightarrow A \rightarrow Y$ is open, whereas the path $A \rightarrow Y \leftarrow L$ is blocked because two arrowheads on the path collide at Y .
2. Any path that contains a non-collider that has been conditioned on is blocked. e.g., the path between A and Y is blocked after conditioning on B.
3. A collider that has been conditioned on does not block a path. e.g., the path between A and Y is open after conditioning on L.
4. A collider that has a descendant that has been conditioned on does not block a path. e.g., the path between A and Y is open after conditioning on C, a descendant of the collider L

Block

Summary:

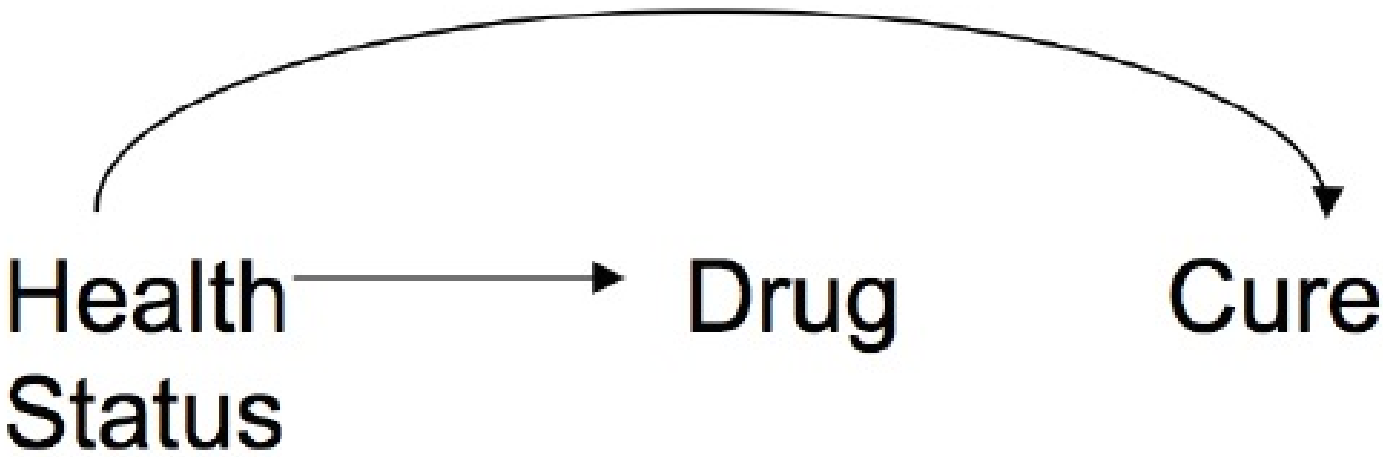
A path is blocked if and only if it contains a non-collider that has been conditioned on, or it contains a collider that has not been conditioned on, and has no descendants that have been conditioned on.

Independence: If all paths between X and Y are blocked conditional on Z, then X and Y are conditionally independent conditional on Z

D-separation

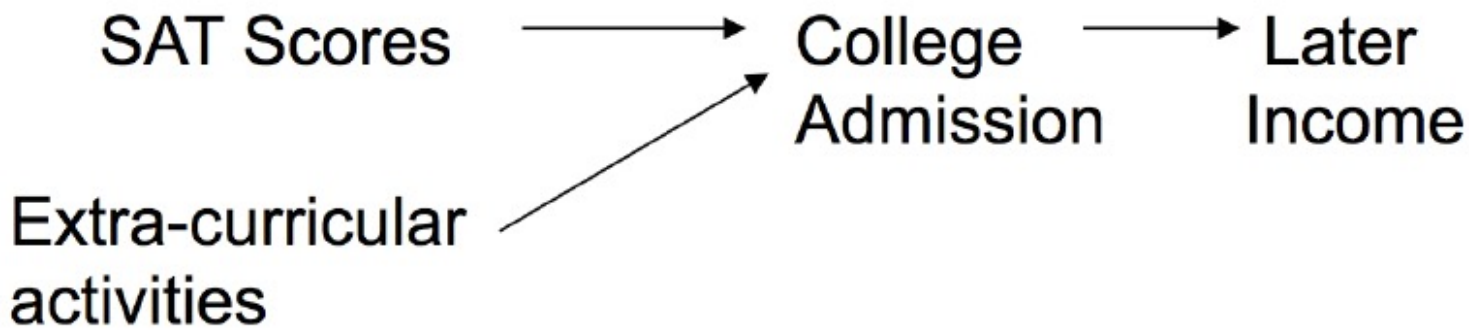
- Two variables are d-separated if all paths between them are blocked (otherwise they are d-connected).
- Two sets of variables are d-separated if each variable in the first set is d-separated from every variable in the second set

Case study: DAGs



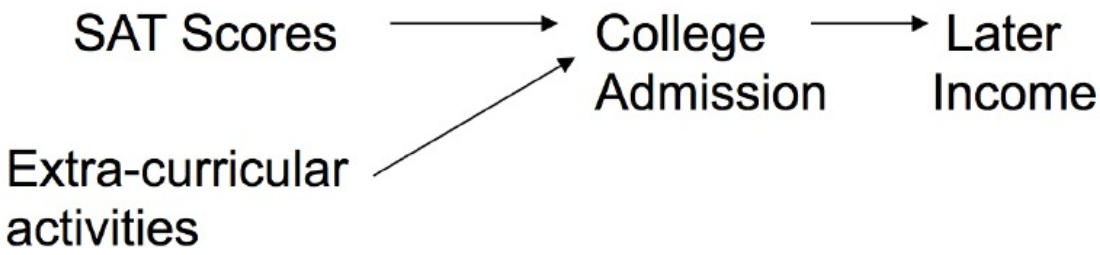
Suppose we are interested in the causal effect of "drug" on "cure", what is "health status"?

Case study: DAGs



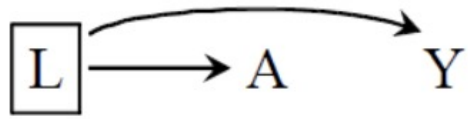
Suppose we are interested in the causal effect of "Extra-curricular activities" on "SAT scores", what is "college admission" and "later income".

Case study: DAGs



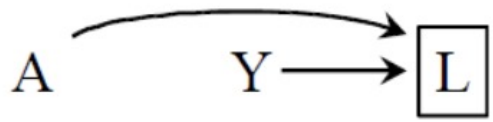
- "SAT scores' is independent of "Extra-curricular activities"
- however, "SAT scores' is not independent of "Extra-curricular activities" conditional on "college admission" or "later income".

Case study: block



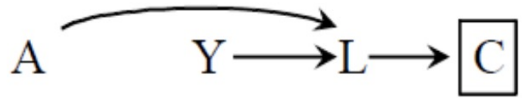
- Conditioning on the common cause L of A and Y along the "backdoor" path $A - L - Y$ blocks the path and makes A conditionally independent of Y given L
- This is the justification for stratification on confounder to adjust for confounding.

Case study: block



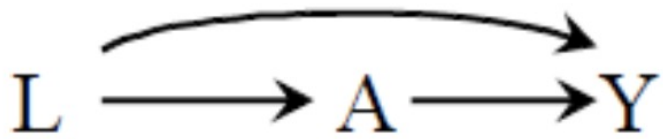
- Conditioning on the collider L creates an association between Y and A, so that while A and Y are marginally independent, they are conditionally dependent given L
- This is known as collider selection bias

Case study: block



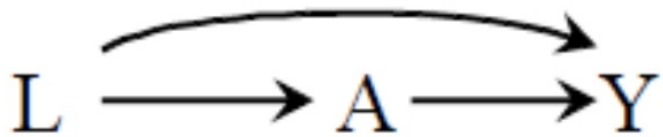
- Suppose the previous DAG is extended as above and C indicates selection into a case-control study
- Graph theory implies that conditioning on a variable C affected by a collider L also opens the path $A \rightarrow L \leftarrow Y$
- C is a common effect of A and Y
- This path is blocked in the absence of conditioning on either the collider L or its consequence C
- But analysis of data implicitly conditions on C

Case study: D-separation



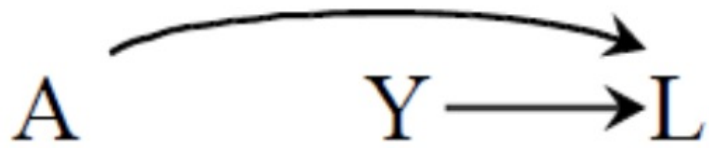
A and L?

Case study: D-separation



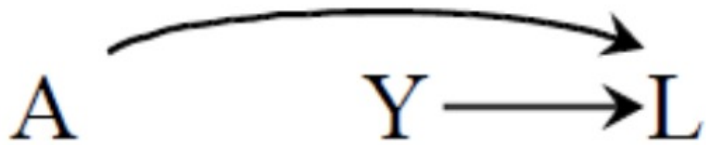
A and L are not marginally d-separated because there is one open path between them ($L \rightarrow A$), despite the other path ($A \rightarrow Y \leftarrow L$) being blocked by the collider Y

Case study: D-separation



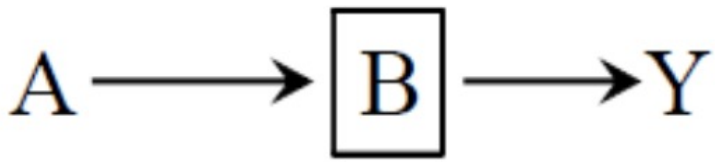
A and Y?

Case study: D-separation



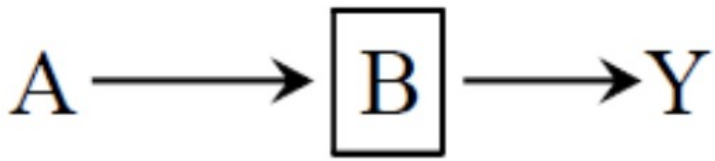
A and Y are marginally d-separated because the only path between them is blocked by the collider L

Case study: D-separation



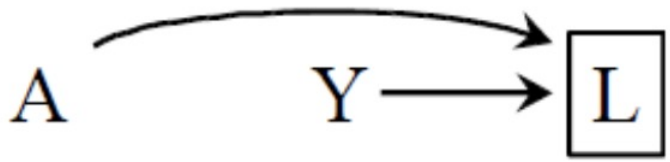
A and Y?

Case study: D-separation



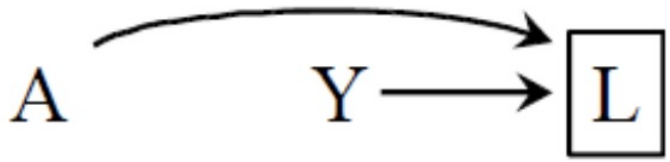
A is conditionally d-separated from Y given (conditional on) B

Case study: D-separation



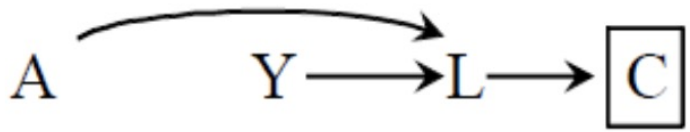
A and Y?

Case study: D-separation



A is conditionally d-connected with Y given (conditional on) L

Case study: D-separation



A and Y?

Discovery of causal structure

Until now, we use causal DAGs to guide the data analysis. Can we go the opposite direction? Can we learn the causal structure by conducting data analyses without making assumptions about the causal structure?

Causal discovery requires that we assume faithfulness so that statistical independence in the observed data distribution imply missing causal arrows on the DAG.

Even with faithfulness assumption, causal discovery is often impossible

Discovery of causal structure

Suppose that we find a strong association between two variables B and C in our data

- B causes C ($B \rightarrow C$)
- C causes B ($C \rightarrow B$)
- B and C share an unmeasured cause $B \leftarrow U \rightarrow C$
- B and C have an unobserved common effect U that has been conditioned on $B \rightarrow U \leftarrow C$
- .....

Discovery of causal structure

Suppose we have an infinite amount of data on 3 variables Z, A, Y and we know that their time sequence is Z first, A second, and Y last. Our data analysis finds that all 3 variables are marginally associated with each other, and that the only conditional independence that holds is $Z \perp Y \mid A$.

- $Z \rightarrow A \rightarrow Y$
- with perhaps a common cause U of Z and A in addition to (or in place of) the arrow from Z to A
- no unmeasured common cause of A and Y exists

Effect modification in DAG

Heart transplant A was randomly assigned in an experiment to identify the average causal effect of A on death Y.

Investigators suspect that the causal effect of heart transplant varies by the quality of medical care offered in each hospital participating in the study ($V = 1$ high quality of care, $V = 0$ normal quality of care)

How to represent the effect modification of V on the causal effect of A to Y using DAG?

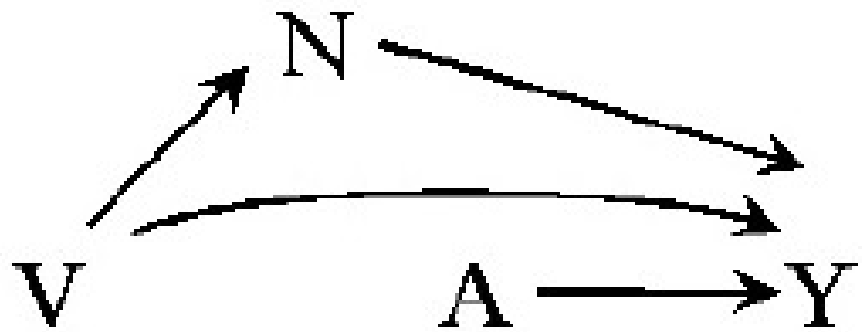
Effect modification in DAG

- arrow from V to Y
- but no arrow into treatment A because A is randomly assigned and thus independent of V
- any caveats?

Effect modification in DAG

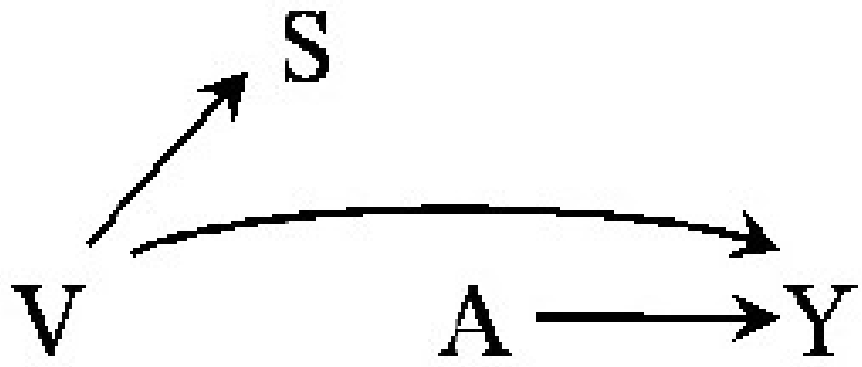
- The common cause does not need to include V becasue V is not a common cause of A and Y
- V is included because we suspect the causal effect of A on Y differ within levels of V
- Other variables measured along the path between "quality of care" V and the outcome Y could also qualify as effect modifiers

Effect modification in DAG



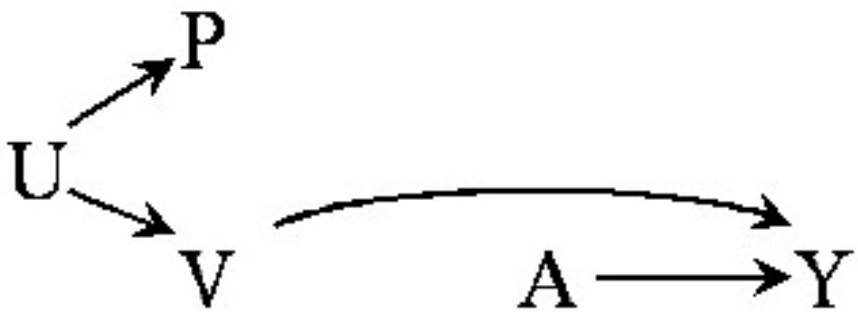
N: "therapy complications"

Effect modification in DAG



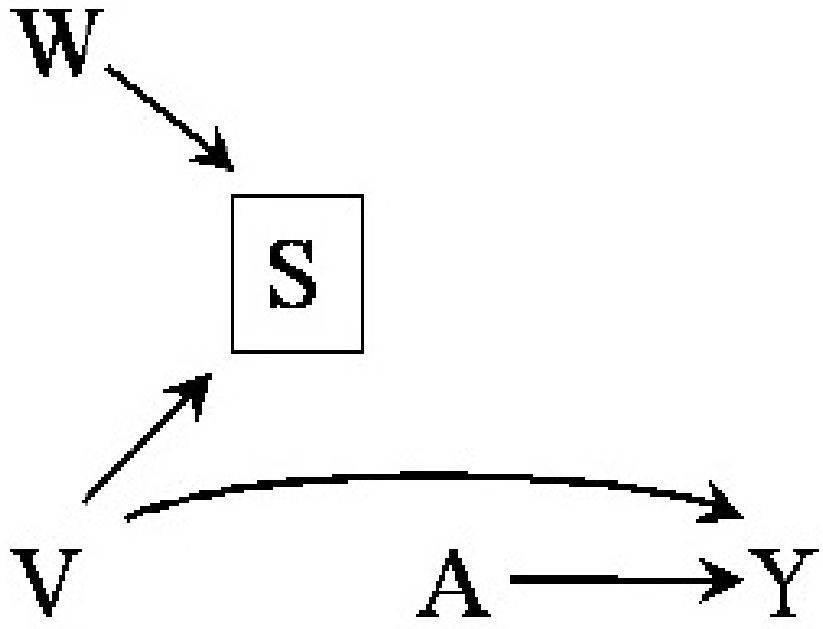
S: "cost of treatment", surrogate effect modifier

Effect modification in DAG



- U: place of residence
- P: passport defined nationality

Effect modification in DAG



- S: cost of treatment
- W: use of bottled mineral water (rather than tap water) for drinking at the hospital

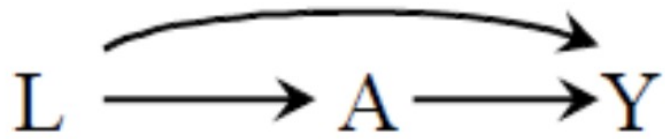
Effect modification in DAG

- The DAG with V could not distinguish which type of effect modification was present
  1. The causal effect of treatment A on mortality Y is in the same direction (i.e., harmful or beneficial) in both stratum $V = 1$ and stratum $V = 0$
  2. The causal effect of treatment A on mortality Y for $V = 1$ is in the opposite directionto stratum $V = 0$
  3. Treatment A has a causal effect on Y in one stratum of V but no causal effect in the other stratum

Effect modification in DAG

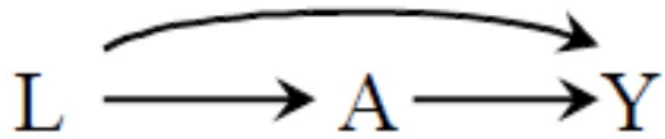
Take home message: effect modification is hard to represent using DAG!

Backdoor criterion



- Confounding bias arises when the treatment and outcome in view share a common cause
- The path $A \leftarrow L \rightarrow Y$ that links A and Y through their common cause L is called a **backdoor path**
- Thus if A is not causally related to Y, i.e. arrow A were deleted, A and Y would be marginally d-connected due to the backdoor path

Backdoor criterion



- Confounding bias arises when the treatment and outcome in view share a common cause
- The path $A \leftarrow L \rightarrow Y$ that links A and Y through their common cause L is called a **backdoor path**
- Thus if A is not causally related to Y, i.e. arrow A were deleted, A and Y would be marginally d-connected due to the backdoor path

Backdoor criterion

- Randomization eliminates all backdoor paths by severing the association between A and L
- More generally, in the absence of randomization, the backdoor criterion states that the treatment effect is identified if one has observed enough variables to block all backdoor paths, that is if treatment and outcome are d-separated given the measured covariates in a graph in which the arrow out of treatment are removed

Backdoor criterion

This criterion answers three questions:

1. does confounding exist?
2. can confounding be eliminated?
3. what variables are necessary to eliminate the confounding?

Crucially it can be used to decide whether one has measured a sufficient set of "confounders" to block all backdoor paths and therefore to adjust for confounding

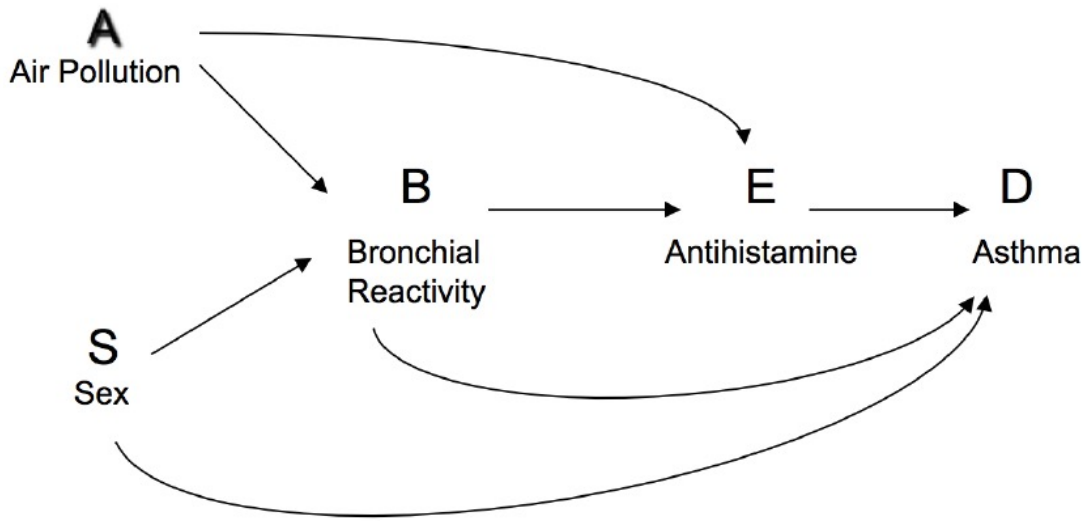
Backdoor criterion

A confounder was traditionally defined as any variable that meets the following three conditions:

1. It is associated with the treatment
2. It is associated with the outcome conditional on the treatment
3. It does not lie on a causal pathway between treatment and outcome

A collider variable meets the same conditions!

Directed Acyclic Graphs (DAGs)

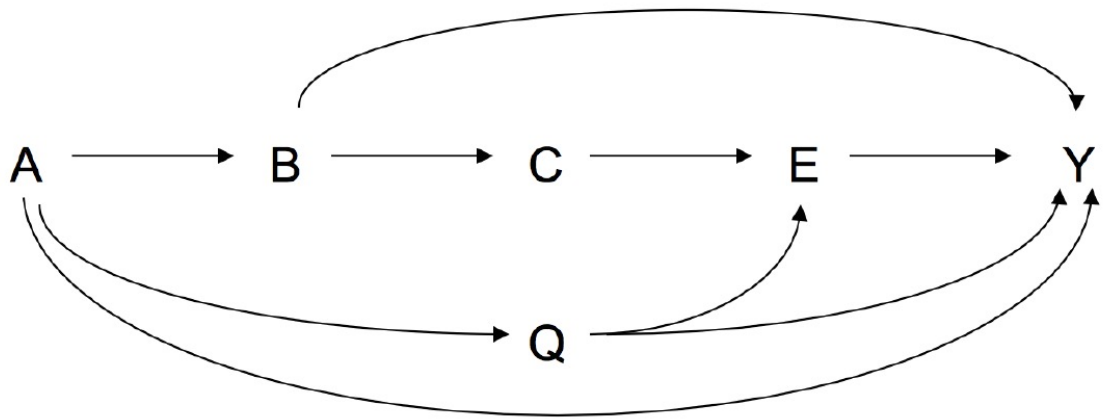


A causal DAG is defined by non-parametric structural equations: $V = f_V(pa_V, \epsilon_V)$ with all V independent (Pearl, 1995).

- $D = f_D(E, B, S, \epsilon_D)$, $E = f_E(A, B, \epsilon_E)$, $B = f_B(A, S, \epsilon_B)$, $A = f_A(\epsilon_A)$, $S = f_S(\epsilon_S)$

Case study: Backdoor Path Criterion

Backdoor Path Criterion (Pearl 1995): For exposure E and outcome Y , if a set of variables Z is such that no variable in Z is a descendent of E and Z blocks all "back-door paths" from E to Y (i.e. all paths from E to Y with edges into E) then conditioning on Z suffices to control for confounding.

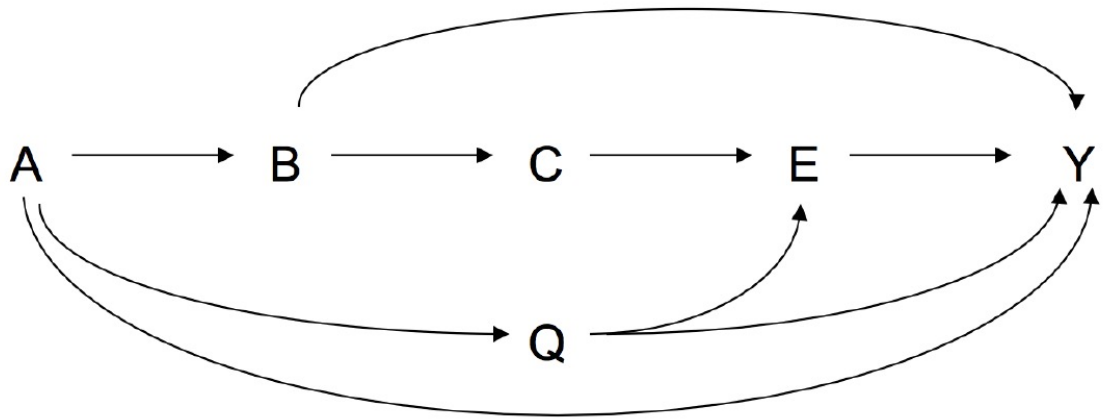


For the causal effect of E on Y :

- Controlling for C and Q suffices?
- Controlling for B and Q suffices?
- Controlling for A and Q suffices?

Case study: Backdoor Path Criterion

Backdoor Path Criterion (Pearl 1995): For exposure E and outcome Y , if a set of variables Z is such that no variable in Z is a descendent of E and Z blocks all "back-door paths" from E to Y (i.e. all paths from E to Y with edges into E) then conditioning on Z suffices to control for confounding.

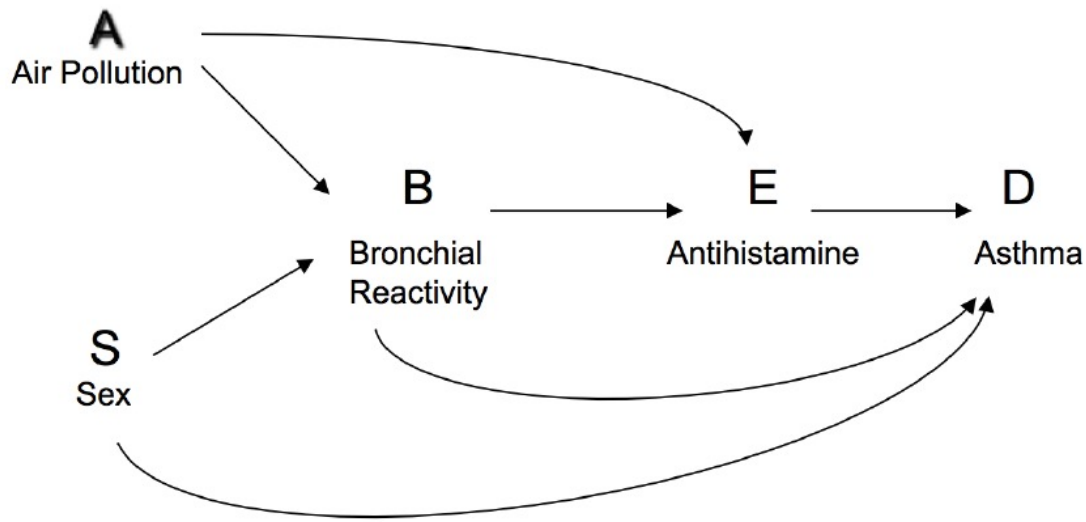


For the causal effect of E on Y :

- Controlling for C and Q suffices
- Controlling for B and Q suffices
- Controlling for A and Q does NOT suffice ($E - C - B - Y$ unblocked)

Case study: Backdoor Path Criterion

Greenland et. al. (1999) gave an example regarding estimating the effect of the use of an antihistamine on asthma

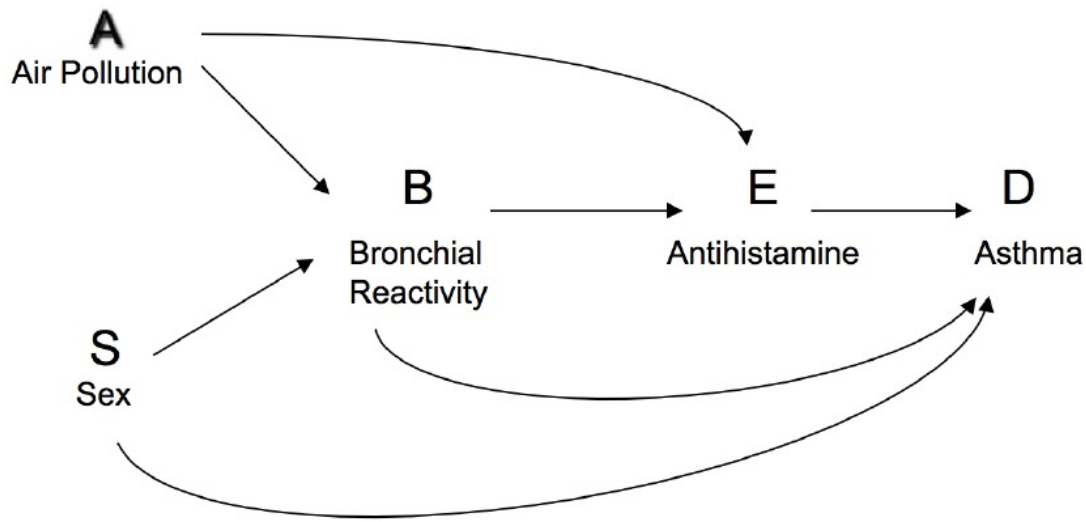


Assumptions in the DAG:

- A affects D only through B and E
- S affects E only through B
- There are no common causes of two variables of the graph that are not on the graph

Case study: Backdoor Path Criterion

Greenland et. al. (1999) gave an example regarding estimating the effect of the use of an antihistamine on asthma

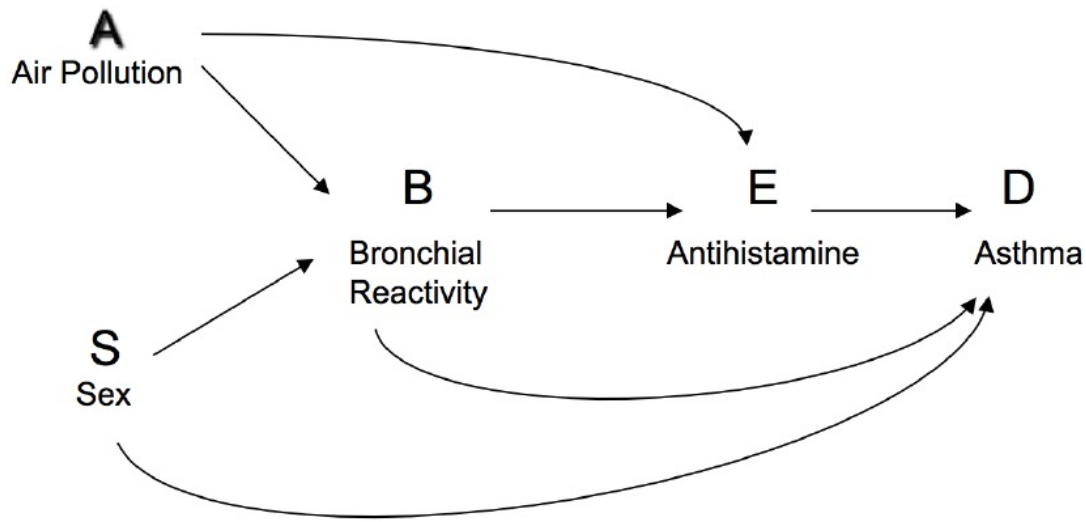


To control for confounding

- condition on A, B and S suffice?
- condition on only A and B suffice?
- condition on only S and B suffice?
- condition on S alone suffice?

Case study: Backdoor Path Criterion

Greenland et. al. (1999) gave an example regarding estimating the effect of the use of an antihistamine on asthma



To control for confounding

- it suffices to condition on A, B and S
- it suffices to condition on only A and B
- it suffices to condition on only S and B
- it does not suffice to condition on S alone

Regression and Causation

Let Y denote a continuous outcome, A exposure and C covariates

$$E[Y \mid A, C] = \beta_0 + \beta_1 A + \beta_2^T C$$

- If the linear regression is correctly specified, but C does not include all the confounders, regression coefficients do not have a causal interpretation but do have an associational interpretation:
- "If we randomly select two individuals from a population and both have the same value of C but the second individual has a value of A one unit higher than the first, then on average, the second individual will have a value of Y that is β_1 units higher."
- Many research studies will appropriately qualify their findings, noting that their results concern association amongst variables and do not necessarily imply causal relationships

Regression and Causation

Regression and Causation: For regression coefficients to have a causal interpretation we need both that

1. The linear regression to be correctly specified
2. All confounders of, e.g., the relationship between treatment A and Y be in the model.

$$E[Y \mid A, C] = \beta_0 + \beta_1 A + \beta_2^T C$$

if $Y^a \perp A \mid C$,

$$E[Y^1 \mid C = c] - E[Y^0 \mid C = c] = E[Y = 1 \mid A = 1, C = c] - E[Y = 0 \mid A$$

i.e., intervening to increase A by one unit will, on average, increase Y by β_1 units.

Regression and Causation

In the absence of interactions between A and C in linear regression:

$$E[Y \mid A, C] = \beta_0 + \beta_1 A + \beta_2^T C$$

if $Y^a \perp A \mid C$,

$$E[Y^1 \mid C = c] - E[Y^0 \mid C = c] = E[Y = 1 \mid A = 1, C = c] - E[Y = 0 \mid A$$

Regression and Causation

In the presence of interactions between A and C in linear regression:

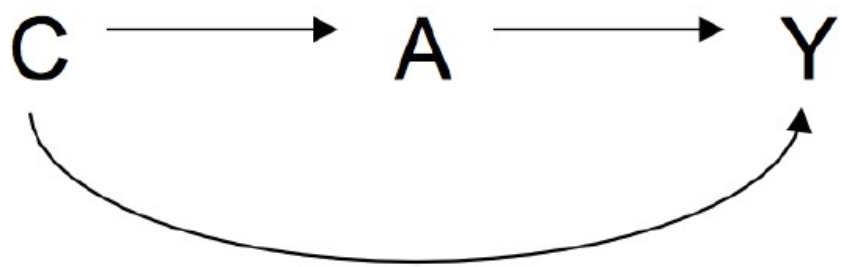
$$E[Y \mid A, C] = \beta_0 + \beta_1 A + \beta_2^T C + \beta_3 A \times C$$

For conditional average causal effects:

$$E[Y^1 \mid C = c] - E[Y^0 \mid C = c] = \beta_1 + \beta_3 \times C$$

What are the implications for inference? The standard error of the effect needs to account for the uncertainty in the estimation of the covariate distributions! How can we account for it? (bootstrap!)

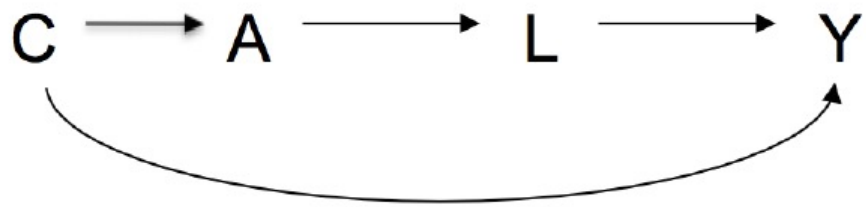
Causal Inference Principle I



- Suppose we wish to estimate the total effect of A on Y
- **Causal Inference Principle I**: If C is a common cause of A and Y, then we should control for C

- If we do not control for C, then the association we observe between A and Y may not be due to the causal effect of A on Y but rather due to the association between A and Y induced by C

Causal Inference Principle II



- Suppose we wish to estimate the total effect of A on Y
- **Causal Inference Principle II**: If there is an intermediate variable L between A and Y, we should not control for it.

- If we do control for L, then some of the association between A and Y due to the causal effect of A and Y may be blocked by controlling for L.

Summary

- Association is not causation and investigators need to be aware under which assumptions statistical analyses yield causal interpretation
- Potential outcomes framework and DAGs help formalizing definition of causal effects, clarifying assumptions, and reason on whether such assumptions are met
- Adjustment for confounding and correct model specification is key to ensure causal interpretation
- Standard regression approaches are typically used for explanatory modeling in the context of fixed time exposures