Using Causal Graphs In Epidemiological Research

Dr David A. Stephens

Department of Mathematics & Statistics

david.stephens@mcgill.ca
www.math.mcgill.ca/dstephens/SISCER2024/

Module Content

1. Probability distributions for observed data.
2. Independence and conditional independence.
3. Graphical modelling.
4. Cause and effect: structural models.
5. Paths and how to block them.
6. d-separation.

Key Challenge in Epidemiological Research

Key objective: causal conclusions from observational data

- Experimental studies:
  - Treatment assigned by the researcher, independent of confounding factors;
  - Causal statements possible.
- Observational studies:
  - Treatment assignment dependent on confounding factors;
  - Causal statements not possible ?

Causal inference

The objective of *causal inference* is to quantify the effect of an *intervention*: in a multi-variable system

- suppose we are able to *manipulate* (i.e. alter the value of) one of the variables separately from all other variables;
- we wish to report the impact of that manipulation on one or more of the other variables.

In many scientific enterprises, this is a primary objective.

Data

We will collect data

$$\{(x_i, y_i, z_i), i = 1, \ldots, n\}$$

which are observed values of the variables X, Y and Z.

- X – predictors, covariates, *confounders*
- Y – *outcome*, response
- Z – *treatment*, exposure

Causal Effects and Counterfactual Outcomes

The *causal effect* of Z on Y is the amount to which an *intervention* to change Z from $\mathtt{z}_0$ to $\mathtt{z}_1$ modifies Y.

- The *potential* or *counterfactual* outcome is denoted $Y(\mathtt{z})$, and is the outcome after an *intervention* to set $Z = \mathtt{z}$.

  For example, if $Z \in \{0, 1\}$, then

  $$Y(0): \text{ outcome if intervention sets } \mathtt{z} = 0 \text{ ('Untreated')}$$

  $$Y(1): \text{ outcome if intervention sets } \mathtt{z} = 1 \text{ ('Treated')}$$

- The *observed* outcome Y is then

$$Y = (1 - Z)Y(0) + ZY(1) = \begin{cases} Y(0) & Z = 0 \\ Y(1) & Z = 1 \end{cases}.$$

Causal Effects and Counterfactual Outcomes

Hypothetical data generating mechanism:

- individual brings their characteristics X;
- for each z, the outcome $Y(\mathsf{z})$ is determined by X;
- for *observed* treatment $Z = z$, we observe $Y = Y(z)$.

Experimental studies

In an *experimental study* precisely the right kind of ‘intervention’ to study causal contrasts is made:

- we randomly assign Z, *independently* of X;
- we *compare* the outcomes in the different groups indexed by different Z values.

Example: randomized controlled trials.

Observational studies

In an *observational study* we do *not* intervene to assign treatments, we observe it as part of the data collection process.

- we *cannot* treat Z as if it were independent of X;
- groups with different Z may have *different* distributions of X, so these groups are *not directly comparable*.

To study cause and effect, we need to have an understanding of the *probabilistic* relationship between all the variables we observe.

To gain statistical insights, we need to build probability models.

Observational studies

Example: HIV Study

- C – CD4 count (continuous)
- D – assignment of new HIV drug (binary)
- H – underlying HIV severity (binary)
- S – symptoms (binary)
- U – follow up indicator (binary, $U = 0$ implies loss to follow-up)

We wish to examine the impact of the new drug on CD4 count:

Does intervening to change D affect outcome C ?

We need to describe how these variables vary jointly in the study.

Probabilistic models

A *joint* probability distribution

$$f_{X,Y,Z}(x, y, z)$$

describes how the data are generated. This model specifies

- the *marginal* distributions

$$f_X(x) \qquad f_Y(y) \qquad f_Z(z)$$

  that describe how the variables behave individually,

- the *conditional* distributions such as

$$f_{X|Y}(x|y) \quad f_{X|Z}(x|z) \quad f_{Y|X}(y|x) \quad f_{Y|X,Z}(y|x,z) \quad f_{Y,Z|X}(y,z|x)$$

  etc. that describe how the variables behave when one or more variable is *fixed*

Probabilistic models

We have the possible decompositions

$$f_{X,Y,Z}(x, y, z) = f_X(x) f_{Z|X}(z|x) f_{Y|X,Z}(y|x, z)$$

$$f_{X,Y,Z}(x, y, z) = f_Z(z) f_{Y|Z}(y|z) f_{X|Y,Z}(x|y, z)$$

and so on, for any ordering of the variables.

We can *always* consider this kind of sequential decomposition, which is termed a *chain rule factorization*.

Independence

Two random variables X, Z are *independent*

$$X \perp\!\!\!\perp Z$$

if and only if, for all values (x, z),

$$\begin{aligned} f_{X,Z}(x, z) &= f_X(x) f_Z(z) \\ f_{Z|X}(z|x) &= f_Z(z) \\ f_{X|Z}(x|z) &= f_X(x) \end{aligned}$$

i.e. knowledge of X does not influence our assessment of Z.

Independence

We can consider *conditional independence*: say

$$Y \perp\!\!\!\perp Z \mid X$$

if and only if, *for all* (x, z, y)

$$f_{Y,Z|X}(y, z|x) = f_{Z|X}(z|x) f_{Y|X}(y|x)$$

i.e. knowledge of Y does not influence our assessment of Z if we *fix* $X = x$.

Example

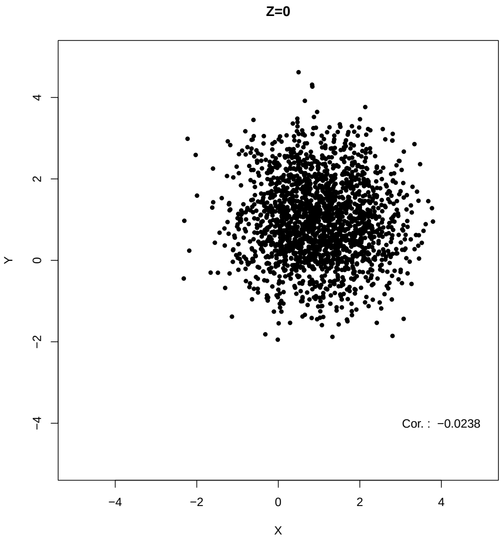
Three variables

- X and Y vary continuously,
- Z is binary.

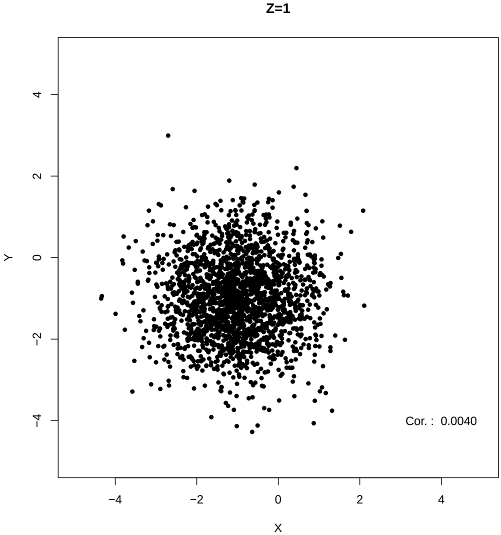
We can study the distribution of the data for X and Y

- for each level of Z separately,
- pooled over Z levels.

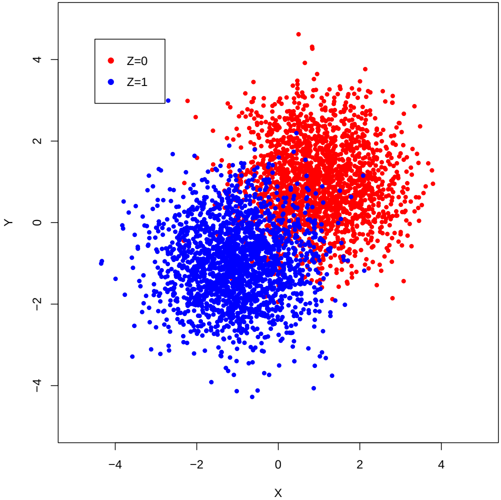
Example



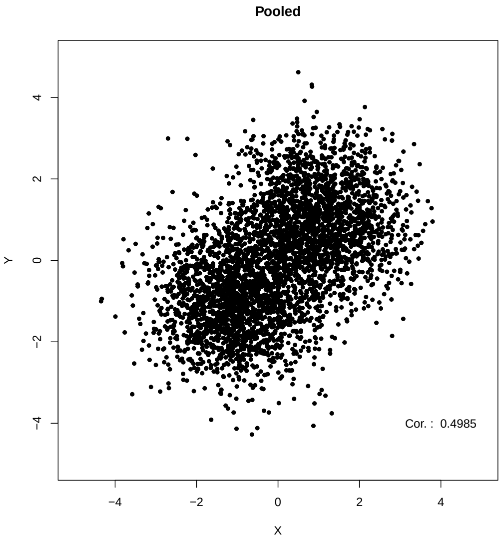
Example



Example



Example



Example

We see that

- for $Z = 0$ and $Z = 1$ separately, X and Y are *uncorrelated*;
- overall X and Y are *positively correlated*.

Thus X and Y are

conditionally unrelated given Z

but are

unconditionally related.

This illustrates that conditioning can remove (or *block*) dependence.

Causal Graphs

We have a chain rule factorization

$$f_{X,Y,Z}(x, y, z) = f_X(x) f_{Y|X}(y|x) f_{Z|X,Y}(z|x, y).$$

We might then *assume* the *conditional independence*

$$Z \perp\!\!\!\perp Y | X$$

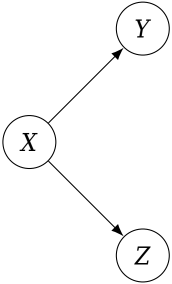
so that

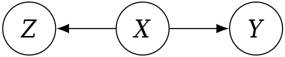
$$f_{Z|X,Y}(z|x, y) = f_{Z|X}(z|x)$$

and so

$$f_{X,Y,Z}(x, y, z) = f_X(x) f_{Y|X}(y|x) f_{Z|X}(z|x)$$

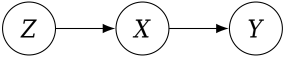
We can depict the conditional independence using a graph:





This type of graph is sometimes called a *fork*.

The other common type of graph component is a *chain*



which implies the factorization

$$f_Z(z)f_{X|Z}(x|z)f_{Y|X}(y|x)$$

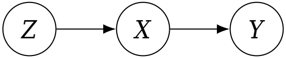
and the conditional independence

$$Y \perp\!\!\!\perp Z|X$$

That is, there are two ways the conditional independence

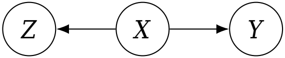
$$Y \perp\!\!\!\perp Z | X$$

could be represented



Chain

$$f_Z(z) f_{X|Z}(x|z) f_{Y|X}(y|x)$$



Fork

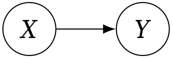
$$f_X(x) f_{Z|X}(z|x) f_{Y|X}(y|x)$$

Causal Graphs

- Nodes (X), (Y), (Z) denote the variables;
- Edges with *arrows* indicate the nature of dependence in the chain rule factorization;
- *Directed* arrows specify the conditional independence assumptions;

Causal Graphs

- Nodes without *incoming* edges are *founders*;



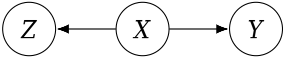
corresponds to

$$f_X(x)f_{Y|X}(y|x)$$

Causal Graphs

- Nodes with only *outgoing* edges act to *block* dependence.

  For example, in

  so that

  $$f_{X,Y,Z}(x,y,z) = f_X(x) f_{Y|X}(y|x) f_{Z|X}(z|x)$$

  it follows that

  $$Z \perp\!\!\!\perp Y | X.$$

  However, it also follows that, in general

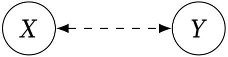
  $$Y \not\!\perp\!\!\!\perp Z$$

  (recall the earlier scatterplots)

Causal Graphs

- *Nodes* or *vertices*, $V_1, V_2, \ldots$, represent variables.
- *Edges*, $E_1, E_2, \ldots$, represent dependencies.
- Two nodes are *adjacent* if there is an edge between them.
  - edges can be *directed*, denoted using arrows, or *undirected*;
  - if all edges are directed, the graph is directed.

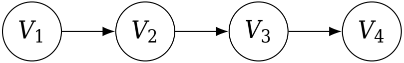
Note: we can use 'bidirected' (edges with an arrow at each end) to indicate general dependence between two variables



although these will be less important in causal settings.

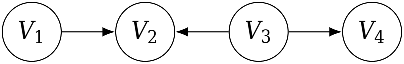
Causal Graphs

- A *path* is a sequence of edges that connects two nodes;
  - a *directed* path is a path where the directions of arrows on edges are obeyed



Directed path from V_1 to V_4

whereas an *undirected* path is a path that is not directed.

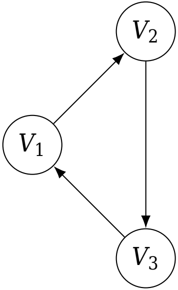


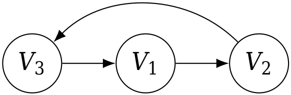
Undirected path from V_1 to V_4

  - two nodes are *connected* if a path exists between them, and *disconnected* otherwise.

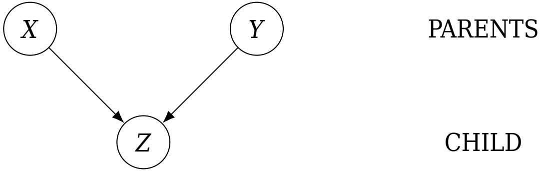
Causal Graphs

In general, a graph may contain *cycles*, that is, directed paths that *start* and *end* at the *same* node.





Directed acyclic graph (DAG): a directed graph with no cycles.



$$f_{X,Y,Z}(x,y,z) = f_X(x) f_Y(y) f_{Z|X,Y}(z|x,y)$$

In this DAG, we have $X \perp\!\!\!\perp Y$:

$$f_{X,Y}(x,y) = f_X(x) f_Y(y)$$

e.g. X and Y represent the scores on two dice rolled independently, Z is the total score

$$Z = X + Y.$$

Causal Graphs

However, conditioning on $Z = z$

$$f_{X,Y|Z}(x, y|z) \neq f_X(x|z) f_Y(y|z)$$

in general. That is,

$$X \perp\!\!\!\perp Y$$

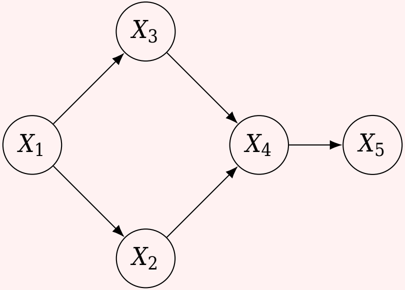
but

$$X \not\!\perp\!\!\!\perp Y \mid Z$$

Conditioning on Z *induces dependence*; the node is termed a *collider*.

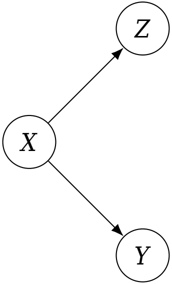
Causal Graphs

Example: Factorization



$$f_{X_1}(x_1) f_{X_2|X_1}(x_2|x_1) f_{X_3|X_1}(x_3|x_1) f_{X_4|X_2,X_3}(x_4|x_2,x_3) f_{X_5|X_4}(x_5|x_4)$$

When we write



what precisely does the symbol $\longrightarrow$ mean ?

Structural models and equations

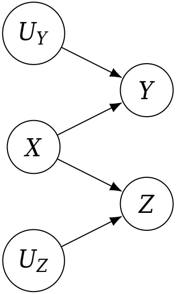
A *structural* interpretation states that we

- generate X independently,
- generate Y and Z independently as functions of the realized X, for example

$$
\begin{aligned}
Y &= 3X \\
Z &= 4X + 9
\end{aligned}
$$

Structural models and equations

X, U_Z, U_Y
independent



$$Y = g_1(X, U_Y)$$

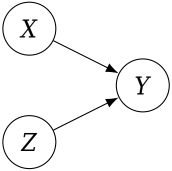
$$Z = g_2(X, U_Z)$$

For example

$$Y = X + U_Y$$

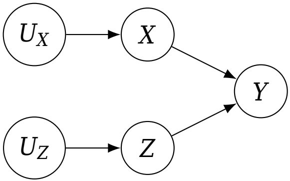
$$Z = X + U_Z$$

Structural models and equations



$$Y = g(X, Z)$$

Fixing $X = x$ and $Z = z$ fixes $Y = g(x, z)$.



$$X = g_1(U_X)$$

$$Z = g_2(U_Z)$$

$$Y = g(X, Z)$$

If we know $X = x$ and $Z = z$, then we do not need to know the values of U_X and U_Z to determine Y. That is

$$Y \perp\!\!\!\perp (U_X, U_Z) \mid (X, Z).$$

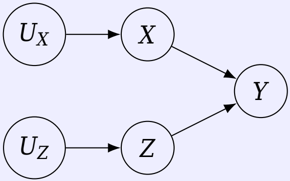
We can interpret *causation* in terms of these functions.

Structural models and equations

- X *causes* Y if it appears in the function, g, that assigns Ys value;
- X *causes* Y if, in the graph representing the joint distribution, there is a *directed path* from X to Y;
- X is a *direct cause* of Y if there is an arrow from X to Y.

Structural models and equations

Note



$$X = g_1(U_X)$$

$$Z = g_2(U_Z)$$

$$Y = g(X, Z)$$

so that

$$Y = g(X, Z) = g(g_1(U_X), g_2(U_Z))$$

so both (X, Z) and (U_X, U_Y) can be interpreted as causes of Y.

- X and Z are direct causes,
- U_X and U_Y are indirect causes.

Structural models and equations

Note

We will proceed by assuming that in a practical setting, the structural relationship and the corresponding causal graph is *known* before any analysis can be carried out.

- Usually in practice this requires expert knowledge;
- Learning the causal graph from data is a hard problem;
- If we obtain all variables simultaneously, it is not possible to learn which of the possible factorizations is the data generating one; for example, we cannot distinguish

$$f_X(x) f_{Y|X}(y|x) \qquad \text{from} \qquad f_Y(y) f_{X|Y}(x|y)$$

i.e. does X cause Y or does Y cause X ?

Structural models and equations

Note

In order for X to cause Y, we must have that X *precedes* Y temporally.

The structural equations form the variables on the left hand side from the variables on the right hand side

$$Y = g(X, Z)$$

that is, we *first* generate X and Z, and *then* generate Y.

That is, there must be a *temporal* ordering.

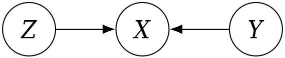
d-separation

To assess whether

$$Y \perp\!\!\!\perp Z \qquad \text{or} \qquad Y \perp\!\!\!\perp Z \mid X$$

for any distribution compatible with the DAG, we must assess whether there is any way for 'information' to 'flow' between Z and Y, maybe once X has been accounted for.

First, recall the *collider* graph



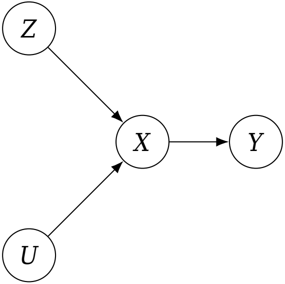
X is a collider on the path between Z and Y. Therefore

$$Y \perp\!\!\!\perp Z \qquad \text{but} \qquad Y \not\perp\!\!\!\perp Z \mid X$$

Note that a *directed path* from one node to another *cannot* contain a collider.

d-separation

The notion of being a collider is *path-specific*: for example



- X is a *collider* on path $Z \to X \to U$
- X is *not a collider* on path $Z \to X \to Y$.

Unconditional d-separation

Consider a general path (directed or undirected) between Z and Y.

The path is *open* (or *unblocked*) if there is *no collider* on the path;

- if there is a collider, the path is *closed* (*blocked*).

Z and Y are *d-separated* if there is *no open path between them*;

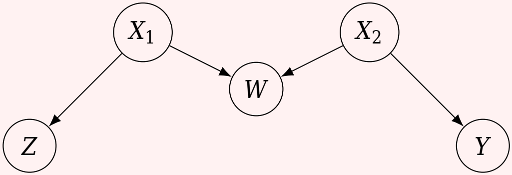
If there is an open path, Z and Y are *d-connected*.

- this path must comprise *chains* or *forks* only

Unconditional d-separation

Example: Diabetes example (Rothman et al. p 188)

- X_1 family income
- X_2 genetic risk
- W parental diabetes
- Z low educational attainment
- Y diabetes of subject



Unconditional d-separation

Example: Diabetes example (Rothman et al. p 188)

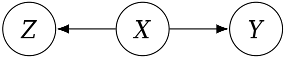
Z and Y are d-separated; there is one path between Z and Y, but it is blocked by the collider W.

$$f_{X_1}(x_1)f_{X_2}(x_2)f_{W|X_1,X_2}(w|x_1, x_2)f_{Z|X_1}(z|x_1)f_{Y|X_2}(y|x_2)$$

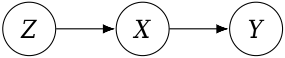
and Z and Y are *independent*.

Conditional d-separation

For a *non-collider* X: *conditioning* on X:



$$Z \perp\!\!\!\perp Y \mid X$$

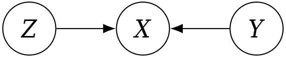


$$Z \perp\!\!\!\perp Y \mid X$$

Conditioning *blocks* the path.

Conditional d-separation

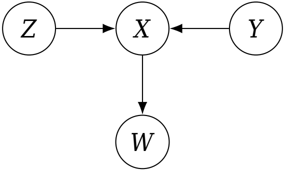
For a *collider* X: conditioning on X *opens* the path



$Z \not\perp\!\!\!\perp Y \mid X$

Conditional d-separation

Consider the following DAG:



and conditioning on *a descendant*, W, of X:

$$f_{Z,Y,W}(x,y,w) = f_Z(z)f_Y(y)\int f_{X|Z,Y}(x|z,y)f_{W|X}(w|x)\,dx$$

$$= f_Z(z)f_Y(y)f_{W|Z,Y}(w|z,y)$$

Conditional d-separation

Therefore we have that

$$Z \not\perp\!\!\!\perp Y \mid W$$

and so W *is a collider* in the reduced graph.

Conditional d-separation

Therefore

(i) conditioning on a *non-collider* X *blocks* the path at X;

(ii) conditioning on a *collider* X or a *descendant* W of X *opens* the path at X;

Conditional d-separation

Consider two nodes X and Y with possibly several open paths connecting them. Suppose S is a set of variables.

- S *blocks* a path if, after conditioning on S, the path is *closed*;
- S *unblocks* a path if after conditioning the path is *open*;
- If S blocks *every path*, then X and Y are *d-separated* by S.

Conditional d-separation

- If S d-separates X and Y, then

$$X \perp\!\!\!\perp Y \mid S,$$

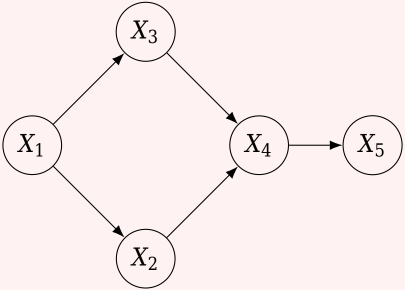
so that

$$f_{X|Y,S}(x|y,s) \equiv f_{X|S}(x|s) \quad \forall (x,y,s).$$

X and Y are *conditionally independent* given S.

Conditional d-separation

Example:

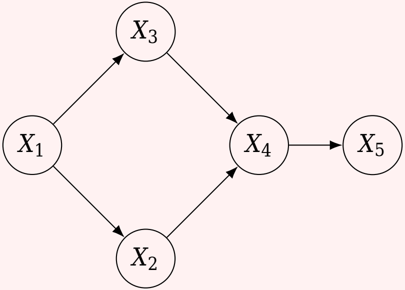


$\{X_2\}$ and $\{X_3\}$ are d-separated by $\{X_1\}$, and $X_2 \perp\!\!\!\perp X_3 \mid X_1$.

- there are two paths between X_2 and X_3;
  - $X_2 \to X_1 \to X_3$: blocked by conditioning on X_1.
  - $X_2 \to X_4 \to X_3$: blocked by the collider at X_4, and $X_4 \notin \{X_1\}$.

Conditional d-separation

Example:



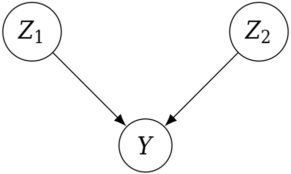
$\{X_2\}$ and $\{X_3\}$ are *not* d-separated by $\{X_1, X_5\}$:

- $X_2 \not\perp\!\!\!\perp X_3 \mid (X_1, X_5)$.
- X_5 is a descendant of collider X_4;

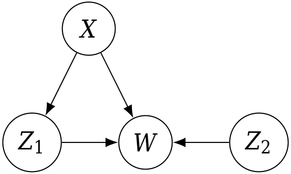
Selection bias

Conditioning on the common effect of two causes renders the two causes dependent;

- this is known as *selection bias* or *Berkson bias*
- it is the effect we observe in the collider graph



Selection bias

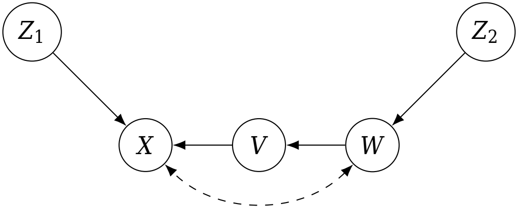


Here $Z_1 \perp\!\!\!\perp Z_2$: there are two paths to consider

- $Z_1 \to X \to W \to Z_2$
- $Z_1 \to W \to Z_2$

both blocked by collider W. Therefore $Z_1 \not\!\perp\!\!\!\perp Z_2 \mid \{W\}$.

Selection bias



Here $Z_1 \perp\!\!\!\perp Z_2$: there are two paths to consider

- $Z_1 \rightarrow X \rightarrow V \rightarrow W \rightarrow Z_2$ is blocked by the collider X.
- $Z_1 \rightarrow X \rightarrow W \rightarrow Z_2$ is blocked by the colliders X and W.

Therefore $Z_1 \not\perp\!\!\!\perp Z_2 \mid \{X, W\}$.

d-separation implication

If X and Y are d-separated by S then

$$X \perp\!\!\!\perp Y \mid S$$

for *all* distributions compatible with the graph; conversely, if they are not d-separated, then X and Y are *dependent* given S for at least one distribution compatible with the graph.

Interventions

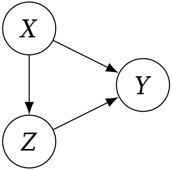
Intervening set the level of Z to z has the effect of

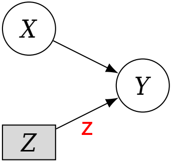
- removing all *incoming* arrows to Z
- switching the marginal for Z to the *degenerate distribution* $f_Z^*(.)$

$$f_Z^*(z) = \mathbb{1}_{\{\mathsf{z}\}}(z) \quad z \in \mathbb{R}.$$

That is, Z takes the value z with probability 1.

In the earlier example

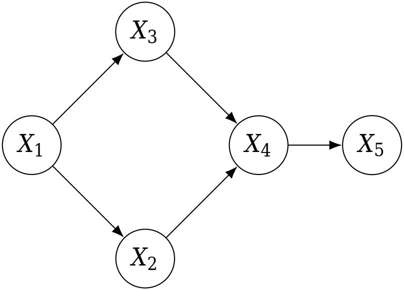




$$f_X(x)f_{Z|X}(z|x)f_{Y|X,Z}(y|x,z) \qquad f_X(x)f_Z^*(\mathsf{z})f_{Y|X,Z}(y|x,\mathsf{z})$$

Interventions

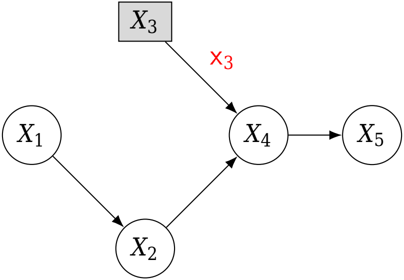
Consider the DAG



$$f_{X_1}(x_1)f_{X_2|X_1}(x_2|x_1)f_{X_3|X_1}(x_3|x_1)f_{X_4|X_2,X_3}(x_4|x_2,x_3)f_{X_5|X_4}(x_5|x_4)$$

Interventions

Suppose we *intervene* to set $X_3 = \mathsf{x}_3$. The relevant DAG is



$$f_{X_1}(x_1) f_{X_2|X_1}(x_2|x_1) f^*_{X_3}(\mathsf{x}_3) f_{X_4|X_2,X_3}(x_4|x_2,\mathsf{x}_3) f_{X_5|X_4}(x_5|x_4)$$

and X_1 is *no longer a cause* of X_3.

Graphical representation of bias

We aim to understand the effect of Z on Y.

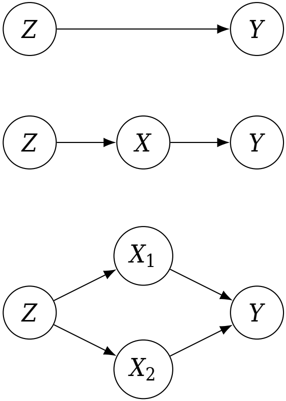
- An *open* undirected path between Z and Y allows for the *association* between Z and Y to be *modified* by the presence of other variables.

  This is known as a *biasing* path.

- By *'association'*, we mean some form of *correlation*.

Graphical representation of bias

- The *association* between Z and Y is *unbiased* for the effect of Z on Y if the only open paths between them are *directed paths*.



Graphical representation of bias

A set of nodes S is *sufficient* to control bias in the association between Z and Y if

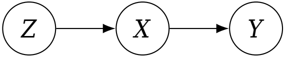
- conditional on S, the *open* paths between Z and Y are precisely the *directed* paths between Z and Y.

S is *minimally sufficient* if it is the smallest sufficient set.

Graphical representation of bias

Note: Conditioning on descendants of Z

(i) *blocks* directed paths

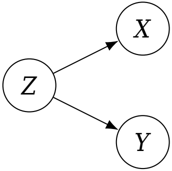


$$Z \perp\!\!\!\perp Y \mid X \quad \text{but} \quad Z \not\!\perp\!\!\!\perp Y$$

(ii) may *unblock* or *create* paths that lead to *biasing* of the effect of Z on Y.

Graphical representation of bias

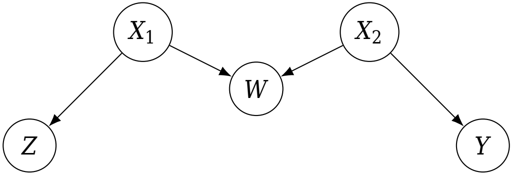
(iii) may be unnecessary in statistical terms: for example



In this graph, conditioning on X will not affect bias.

Graphical representation of bias

Undirected paths from Z to Y are termed *backdoor* paths (relative to Z) if they *start* with an arrow pointing *into* Z.



The only path from Z to Y is a backdoor path; however, it is not open because of the collider W.

Graphical representation of bias

Before conditioning

- *all biasing* paths in a DAG are backdoor paths, and
- all *open* backdoor paths are biasing paths.

To obtain an unbiased estimate of the effect of Z on Y, all backdoor paths between Z and Y must be *blocked*.

Graphical representation of bias

Set S satisfies the *backdoor criterion* with respect to Z and Y if

(i) S *contains no descendant* of Z, and

(ii) there is *no open backdoor path* from Z to Y after *conditioning* on S.

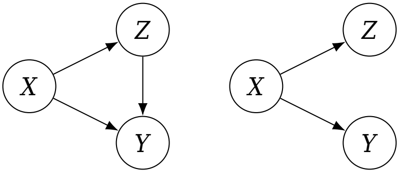
A *confounding path* between Z and Y is

(i) a *biasing* path (that is, an *undirected open path*) that

(ii) *ends* with an arrow into Y.

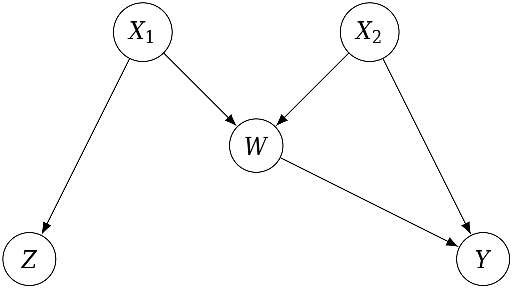
Confounding

Variables on a confounding path are termed *confounders*.



X is a confounder in both cases.

Confounding



W is a collider on the undirected path from Z to Y

$$\text{Path 1: } Z \rightarrow X_1 \rightarrow W \rightarrow X_2 \rightarrow Y$$

and hence this path is blocked.

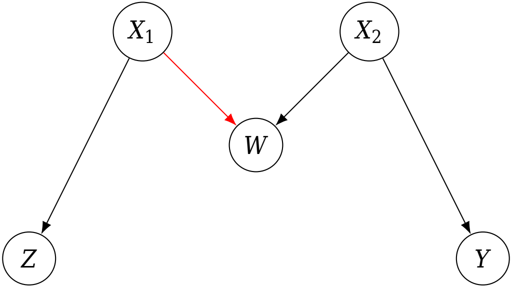
However unconditional on W, the effect of Z on Y is confounded by the backdoor path

$$\text{Path 2: } Z \rightarrow X_1 \rightarrow W \rightarrow Y.$$

Conditioning on W alone opens Path 1, therefore to block both paths, we need to condition on

$$S \equiv \{W, X_2\}.$$

Confounding

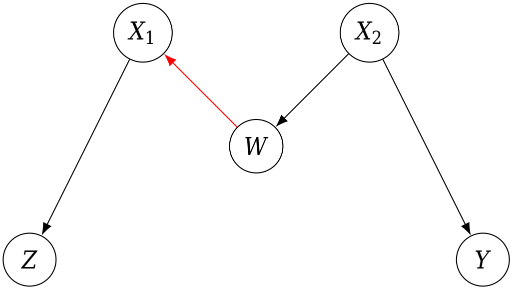


Conditioning on W *opens* the confounding path. Therefore $Z \perp\!\!\!\perp Y$ (as there is no open path between them), but

$$Z \not\perp\!\!\!\perp Y \mid W$$

Further conditioning on either $\{X_1\}$ or $\{X_2\}$ blocks the path.

Confounding



Conditioning on W *blocks* the confounding path. Therefore conditioning on any one of

$$\{X_1\}, \{W\}, \{X_2\}$$

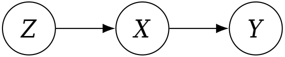
will block the path.

Direct and indirect effects

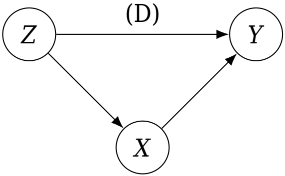
For the effect of Z on Y relative to X:

- *Direct effect:* A direct effect of Z on Y is the effect captured by a *directed* path from Z to Y that does not pass through X.
- *Indirect effect:* An indirect effect of X on Y that is captured by directed paths that pass through X.
    - X is termed an *intermediate* or *mediator* variable.

Direct and indirect effects



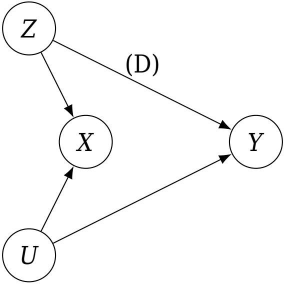
Indirect effect



Direct (D) & Indirect effect

X is a mediator of the indirect effect

Direct and indirect effects



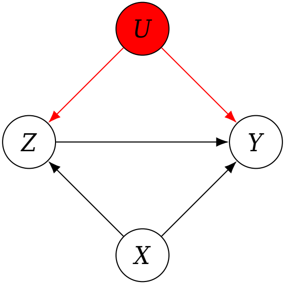
No indirect effect

Direct effect is not confounded

X is a collider, so there is no other open path from Z to Y.

Unmeasured confounding

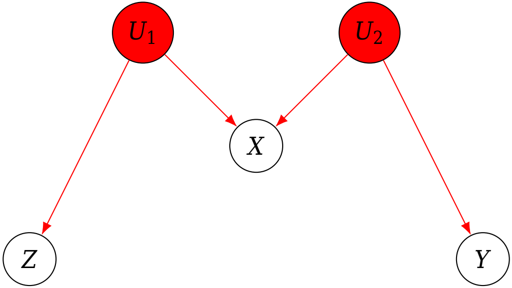
Suppose that in reality there is a further variable U that is a confounder, but is unmeasured in the observed data.



There is a hidden confounding path $Z \to U \to Y$. Conditioning on U is not possible, as we are unaware of its existence.

Unmeasured confounding

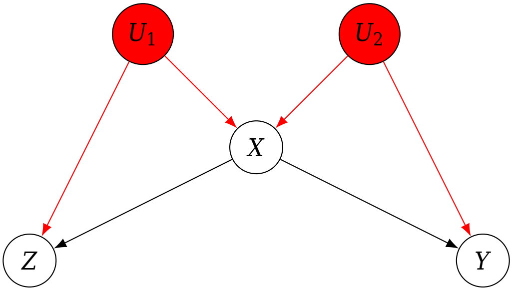
With two unmeasured confounders:



We have that X, Y and Z are *independent*; the (true but hidden) path between Z and Y is blocked at collider X.

Unmeasured confounding

Suppose we condition on X:

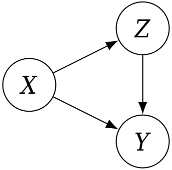


In the modelled DAG, $Y \perp\!\!\!\perp Z \mid X$; however, conditioning on X *opens* the *hidden* path through U_1 and U_2, so there is now an open biasing path.

This is sometimes referred to as the *M-bias* phenomenon.

Implications for Statistical Modelling

We have seen that conditioning on variables can close biasing paths, allowing an unbiased assessment of the causal effect of Z on Y.



The open, undirected path

$$Z \to X \to Y$$

can be blocked by conditioning on X.

Implications for Statistical Modelling

If all the variables are jointly Normally distributed, then this conditioning can be achieved by including X as a predictor in a linear regression model of Y on Z.

That is, we can fit the linear model where

$$\mathbb{E}[Y|X = x, Z = z] = \beta_0 + \beta_1 x + \psi z$$

and estimate the direct effect of Z on Y by estimating ψ.

Implications for Statistical Modelling

Note

Blocking confounding paths (e.g. by conditioning) is not quite the end of the story.

Typically we need to utilize *parametric* inference, and there is usually a requirement that certain parametric models are *correctly specified*.

Summary

In a statistical formulation of a causal inference problem

1. We identify *treatment* Z and *outcome* Y
2. We form the *DAG* representing the relationships between Z and Y which contains other measured variables X.
3. The causal effect of Z on Y flows down *open* and *directed* paths from Z to Y;
   - there may be a *direct* effect if there is an arrow from Z into Y;
   - there may also be *indirect* effects if the directed path passes through *mediating* variables.

Summary

4. If there are *undirected* paths from Z to Y that are open, then these paths may induce *bias* in estimation of the effect of Z on Y.
5. In order to obtain unbiased estimation, the open undirected (biasing) paths must be *blocked*; typically this is done by *conditioning* on variables on those paths.
6. A *collider* node blocks a path; however, conditioning on the collider *opens* the path at that node.