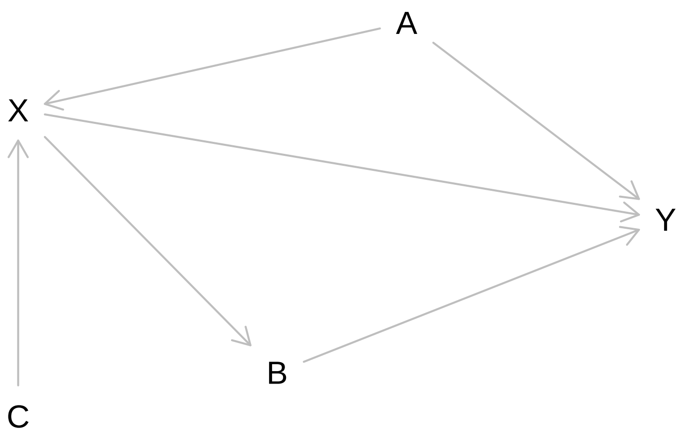
Introduction to DAGs

A *directed acyclic graph* (DAG) is also known as a **causal diagram**. It is a graph that displays the causal relationships between a set of variables. Directed acyclic graphs and the mathematical and statistical tools for understanding them were primarily developed by Judea Pearl and colleagues.

Here is an example DAG:



DAGs consist of *nodes*, which represent variables, and *edges* (arrows) that represent causal relationships.

J. Pearl's working definition of causation is as follows: **variable X is said to cause variable Y if Y depends on X for its value**.

DAGs are useful for several reasons. In fact, you can't use linear regression for explanatory purposes if you don't understand DAGs.

- DAGs help you decide how to specify your model (i.e., what variables to include, and what variables to leave out).
- DAGs clearly communicate your assumptions to the audience.
- DAGs can be used to identify some testable hypothesis from theories.

Basic DAG terminology

node: a variable

edge: a causal relationship, represented by an arrow

exposure: The exposure variable is our *focal predictor* – it is the variable whose causal effect we are attempting to understand. Usually represented by X. The independent variable.

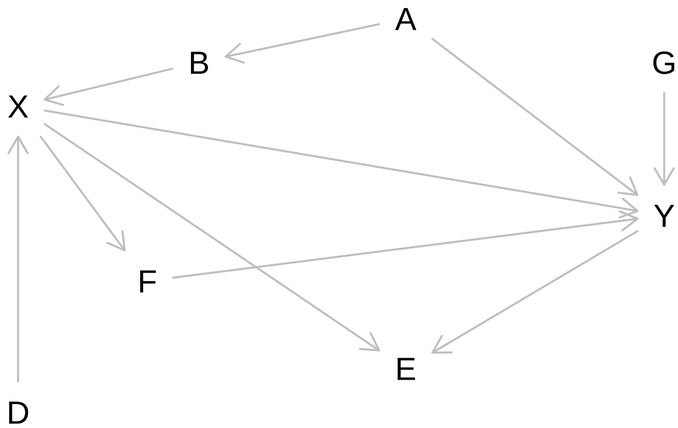
response: the outcome variable, usually noted by Y. The dependent variable.

ancestors: Nodes that are "upstream" from a particular variable. In the above DAG, A and C are ancestors of X.

decendents: Nodes that are "downstream" from a particular variable. In the above DAG, Y and B are descendants of X.

Variable Roles in DAGs

Here's another example DAG.



confounder: A confounder is an ancestor of both the exposure X and the response Y. In the example DAG above, **A** is the only confounder.

mediator: A mediator is a descendant of the exposure variable and an ancestor of the response variable. **F** is a mediator. Some (or all) of the causal effect of the exposure on the outcome is transmitted via the mediator.

proxy confounder: A proxy confounder is the descendant of a confounder and a ancestor of either the exposure or response (but not both, otherwise it would be a confounder). The confounding effect is transmitted through this variable. **B** is a proxy confounder in the DAG.

competing exposure: A competing exposure variable is an ancestor of the response variable Y that is neither an ancestor or nor a descendant of the exposure X. **G** is a competing exposure.

instrument: An instrument is a an ancestor of the exposure variable. It can have no path to the response Y that does not pass through the exposure X. (Otherwise it would be a confounder). **D** is an instrument.

collider: A collider is a descendant of both the exposure X and the response Y. **E** is a collider in the example DAG.

Statistical Vocabulary

independence: A pair of variables, X and Y, are said to be independent if knowledge of one gives no knowledge of the other. Note: independence implies uncorrelated, but uncorrelated does not necessarily imply independence (because they could be dependent but not *linearly* related)

conditioning: To hold a variable constant. This can be accomplished by regression adjustment, sampling, or matching.

conditional independence: Variables X and Y are conditionally independent given Z if knowing X gives no information about Y (or vice versa) after conditioning on Z.

marginal independence: Independence that exists in the absence of conditioning.

dependence: The opposite of independence.

The DAG completely describes the both the marginal and conditional independence for all the variables.

More Vocabulary

Chain: Example: $A \to B \to C \to D \to E$. This is a chain of causation from A to E.

Fork: Example: $A \leftarrow B \leftarrow C \to D \to E$. Variable C forks into two chains, $C \to B \to A$ and $C \to D \to E$.

Spurious path: A connection between an exposure variable and an outcome that induces statistical dependence between them but does not depend on directed paths from X to Y.

***d*-separation**: The process of determining, given the DAG and a set of conditioning variables, whether a pair of variables is statistically independent or dependent.

***d*-connected**: Two nodes (variables) are said to be **d-connected** when an unblocked connecting path exists between them. (The d means *directional*).

***d*-separated**: Two nodes (variables) are said to be **d-separated** when no connecting path exists between them, or when every path connecting them is *blocked*.

blocked: A path is *blocked* when a node along the path is conditioned on.

testable implication: Using data to test for the pattern of conditional and unconditional independences implied by the DAG. This provides a means of testing whether a DAG is plausible causal model for a dataset and also of identifying, to some extent, *how* the model is wrong.

Rules for Decomposing DAGs

Rule 1 (Conditional Independence in Chains) *Two variables, X and Y, are conditionally independent given Z, if there is only one unidirectional path between X and Y and Z is any set of variables that intercepts that path.* (Pearl et al, 2016, p. 39).

Rule 2 (Conditional Independence in Forks) *If a variable X is a common cause of variables Y and Z, and there is only one path between Y and Z, then Y and Z are independent conditional on X.* (p. 40)

Rule 3 (Conditional Independence in Colliders) *If a variable Z is the collision node between two variables X and Y, and there is only one path between X and Y, then X and Y are unconditionally independent but are dependent conditional on Z and any descendents of Z.* (p. 44).

Formal definitions (from Pearl et al., 2016)

***d*-separation** *A path p is blocked by a set of nodes Z if and only if:*

1. *p contains a chain of nodes $A \to B \to C$ or a fork $A \leftarrow B \to C$ such that that middle node B is conditioned on, or*
2. *p contains a collider $A \to B \leftarrow C$ such that neither the collision node B, nor any of its descendents, are conditioned on.* (p. 46)

The Backdoor Criterion *Given an ordered pair of variables (X, Y) in a DAG, a set of variables Z satisfies the backdoor criterion relative to (X, Y) if no node in Z is a descendent of X, and Z blocks every path between X and Y that contains an arrow into Y.* (p. 61).

If a set of conditioning variables Z satisfies the backdoor criterion for X and Y, the the causal effect of X on Y can be computed without manipulating X.

If the set of conditioning variables, Z, satisfies the *backdoor criterion* then:

1. *All spurious paths from X to Y are blocked.*
2. *All directed paths from X to Y are unblocked.*

3. *No new spurious paths are created (by including a collider)* (p. 61).

The Frontdoor Criterion *A set of variables Z is said to satisfy the front door criterion relative to an ordered pair of variables (X, Y) if*

1. *Z intercepts all directed paths from X to Y.*
2. *There is no unblocked path from X to Z.*
3. *All backdoor paths from Z to Y are blocked by X* (p. 69).

If Z satisfies the front-door criterion relative to (X, Y) and if $p(x, z) > 0$, then the causal effect of X on Y is identifiable.

This criterion provides a means of computing causal effects absent manipulation when the backdoor criterion cannot be satisfied. Note that criteria (1) and (2) imply that Z completely mediates the effect of X on Y.