A tutorial on causal inference

Andrea Rotnitzky

Dep. of Economics, Universidad Di Tella, Buenos Aires and Dep. of Biostatistics, Harvard School of Public Health

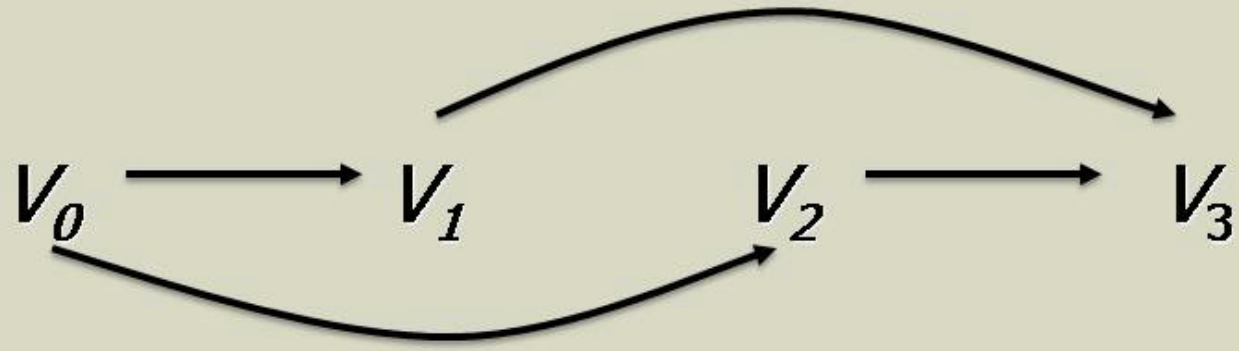
Section I: Directed Acyclic Graphs and Bayesian Networks

- Definition of Directed Acyclic Graphs
- DAG configurations.
- Bayesian networks
- d-separation
- The Markov Factorization Theorem.

DIRECTED ACYCLIC GRAPHS (DAGS)

- A **graph** consists of a set V of vertices (or nodes) and a set E of edges (or links) that connect some pairs of vertices. ..
- A **directed graph** is a graph consisting of directed edges ; i.e. each edge is marked by a single arrowhead.
- A **directed path** in a graph is a sequence edges, each edge pointing to a node from which the next edge emerges.
- A **path** in a graph is a sequence (directed or not) of edges such that each pair of consecutive edges in the sequence share one node.
- A **cycle** is any directed path that starts and ends at the same node.
- A graph that contains no directed cycles is called **acyclic**

DAGS: Directed Acyclic Graphs

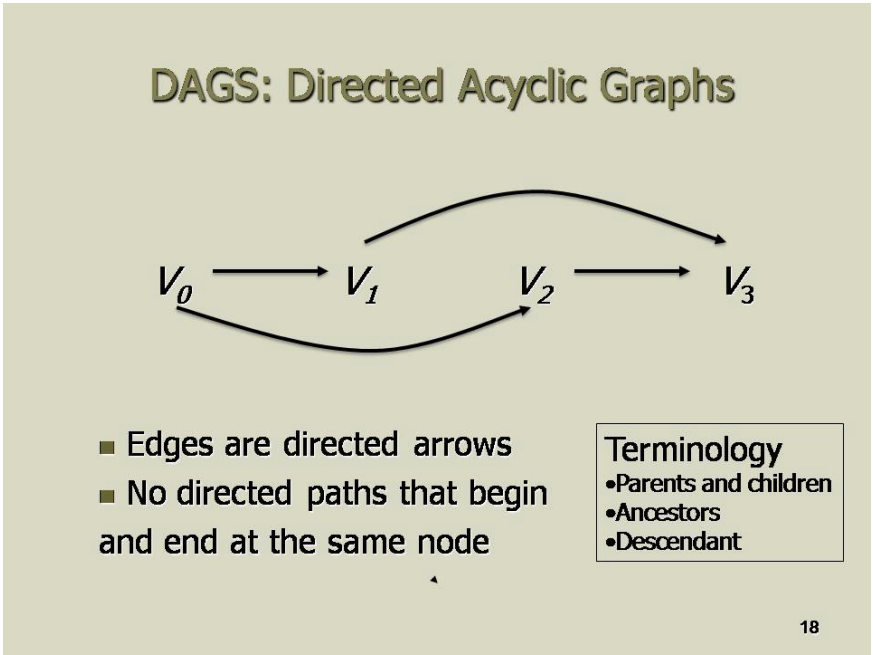


- Edges are directed arrows
- No directed paths that begin and end at the same node

Terminology
- Parents and children
- Ancestors
- Descendant

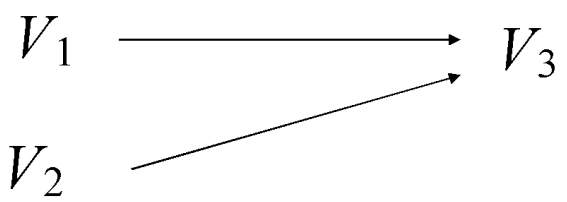
18

- **Definition.** the ordering $(V_1, ..., V_K)$ **agrees with the DAG** iff $\overline{V}_i \equiv \{V_1, ..., V_{i-1}\}$ does not include any descendant of V_i. for each i.
- **Example.**

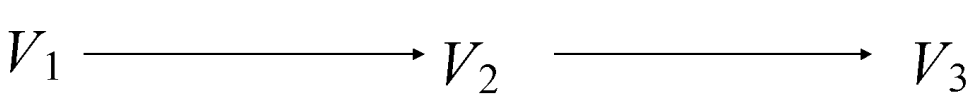


- (V_0, V_1, V_2, V_3) agrees with the DAG
- (V_0, V_2, V_1, V_3) agrees with the DAG,
- (V_1, V_0, V_2, V_3) does not agree with the DAG.

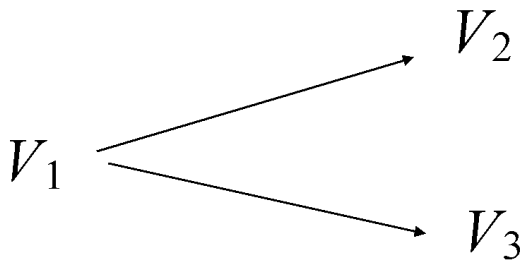
DAG CONFIGURATIONS



COLLIDER, OR INVERTED FORK



CHAIN



FORK

19

What are we aiming for....

- Suppose you know that the law p of $V = \{V_1, ..., V_k\}$ satisfies

$$p(V) = \prod_{i=1}^{k} p(V_i | PA_i) \qquad \boxed{\textbf{Markov Decomposition}} \qquad (1)$$

for some subsets $PA_i \subseteq \{V_1, ..., V_{i-1}\}$.

- Your goal is to determine all conditional independencies

$$X \amalg Y | Z$$

between any three disjoint subsets X, Y and Z of V that are logically implied by Markov decomposition.

- **Notation:** $X \amalg Y | Z$, iff. X and Y are conditionally independent given Z

What are we aiming for....

- We will learn a graphical algorithm to achieve your goal without any calculations!

- **Algorithm:**

  1. Construct the DAG with nodes V and with arrows from each element of PA_i to V_i (for all i)
  2. Are X and Y **d-separated** by the set Z in the DAG?

     1. If yes, conclude that $X \amalg Y | Z$
     2. If not, conclude that $X \amalg Y | Z$ is not logically implied by the Markov decomposition.

- **Disclaimer**: all random vectors are *discrete*, i.e. absolutely continuous with respect to the counting measure
- **Notational remark**. p stands for the mass probability of some random vector. Which vector p is the law for, will be clear from its variables. Thus, for example,
  - $p(v)$ stands for $\Pr(V = v)$
  - $p(y|x)$ stands for $\Pr(Y = y|X = x)$
- $p(V)$ stands for the density of V evaluated at a random value V, etc. Thus, for example,

$$p(V) = \prod_{i=1}^{k} p(V_i|PA_i)$$

is equivalent to

$$\Pr(V = v) = \left\{ \prod_{i=1}^{k} \Pr(V_i = v|PA_i = pa_i) \right\} I_{\{p(\cdot)\}}(v)$$

for all $v \in R^k$

d-separation

- **Definition**: A path is said to be **d-separated, blocked or rendered inactive**, **by a set of nodes** Z if and only if

  1. the path contains a chain $V_i \rightarrow V_m \rightarrow V_j$ or a fork $V_i \leftarrow V_m \rightarrow V_j$ such that the middle node V_m is in Z,

     or

  2. the path contains a collider $V_i \rightarrow V_m \leftarrow V_j$, such that neither V_m *nor its descendants* are in Z.

- **Definition**: A set of nodes Z is said **d-separate** a set of nodes X from another set of nodes Y if and only if Z **blocks every path** from a node in X to a node in Y.

- **Notation:**

$$(X \amalg Y | Z)_G \text{ iff } Z \text{ d-separates } X \text{ from } Y \text{ in } G$$

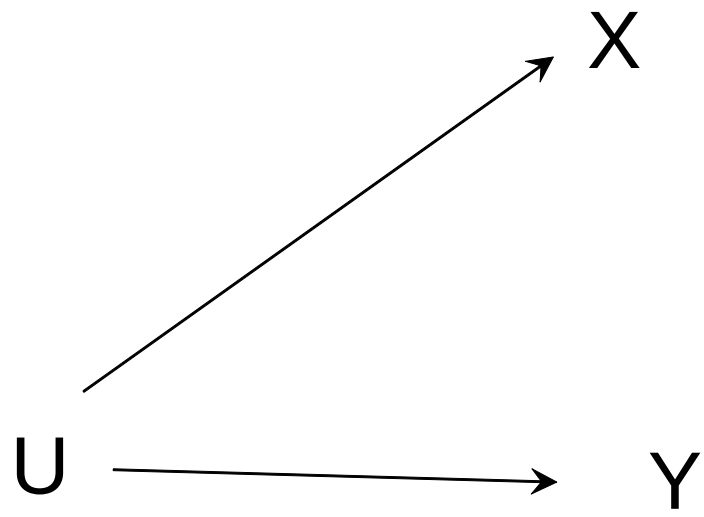
d-separation

- A path is said to be **d-connected by a set of nodes** Z iff it is not d-separated by Z

- **Notational remark:**

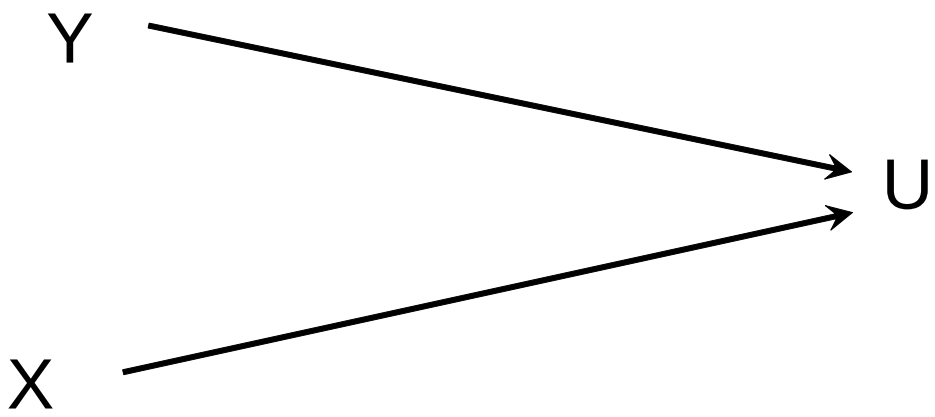
  1. $(X \amalg Y|Z)_G$ means X and Y are **d-separated** by Z

  2. $(X \amalg Y|Z)_P$ means X and Y **are conditionally independent** given Z when they have joint distribution P.



Z = {U} then path between X and Y blocked by Z

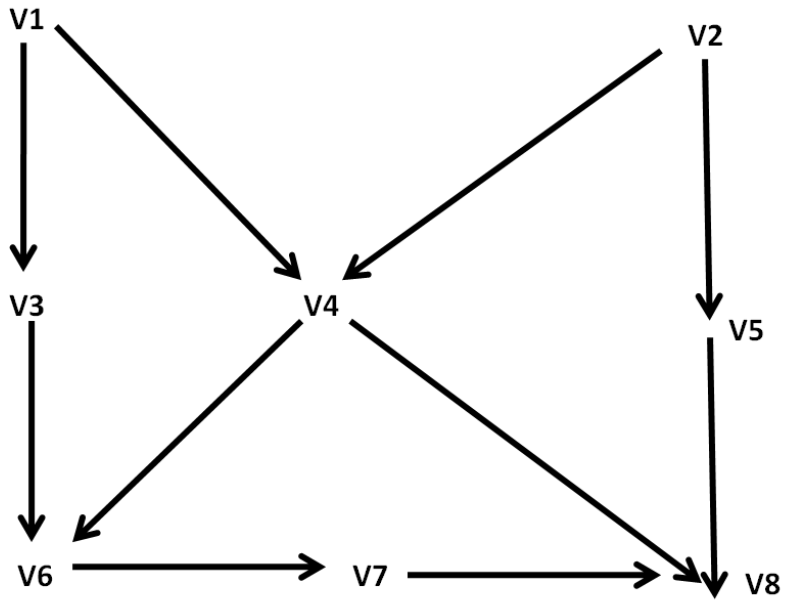
Z= { } then path between X and Y is unblocked by Z



Z = {U} then path between X and Y is **unblocked** by Z

Z= { } then path between X and Y is **blocked** by Z

d-separation and d-connection: more examples



- $(V_6 \amalg V_8 | \{V_7, V_4, V_2\})_G$ and $(V_6 \amalg V_8 | \{V_7, V_4, V_1\})_G$.
- $(V_6 \not\amalg V_8 | \{V_7, V_4\})_G$ because V_4 unblocks the path $V_6, V_3, V_1, V_4, V_2, V_5, V_8$.

The main result

- **Definition:** Given a DAG G with nodes $V = \{V_1, ..., V_k\}$ and a law P of V, we say that G represents P iff

$$p(V) = \prod_{i=1}^{k} p(V_i | PA_i) \tag{2}$$

where PA_i are the parents of V_i on the DAG.

- **Definition**: a DAG and the collection of all $P's$ represented by it is called a **Bayesian Network**

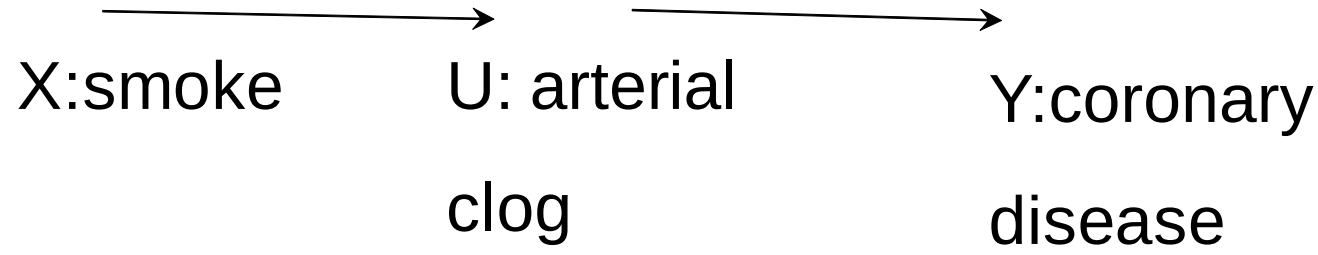
- **Theorem:** *Verma and Pearl (1988) and Geiger (1988).*

Let X, Z and Y be three disjoint sets of nodes in a DAG G. Then

$$(X \amalg Y | Z)_G \Leftrightarrow (X \amalg Y | Z)_P \text{ for all } P \text{ represented by } G$$

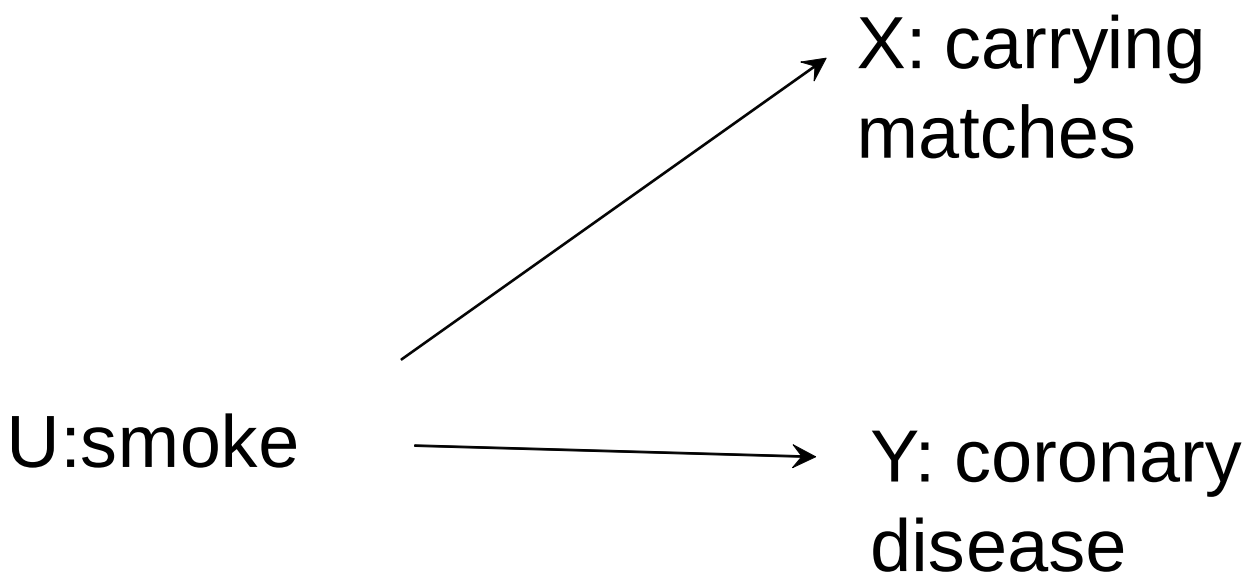
Remarks

- $d-$separation encodes all conditional independencies logically implied by the Markov factorization of any P that is represented by the DAG.
- DAGs carry assumptions through their missing arrows, not through their existing arrows.
- If $(X \not\perp\!\!\!\perp Y|Z)_G$ then there exist at least one law P represented by G such that $(X \not\perp\!\!\!\perp Y|Z)_P$.
  - **Be careful:** $(X \not\perp\!\!\!\perp Y|Z)_G$ does not imply that $(X \not\perp\!\!\!\perp Y|Z)_P$ holds for *all laws* P represented by G.
  - **Example**: a *complete* DAG represents *all laws* P. In complete DAGS no (X, Z, Y) satisfies $d-$separation, yet for some laws $(X \amalg Y|Z)_P$



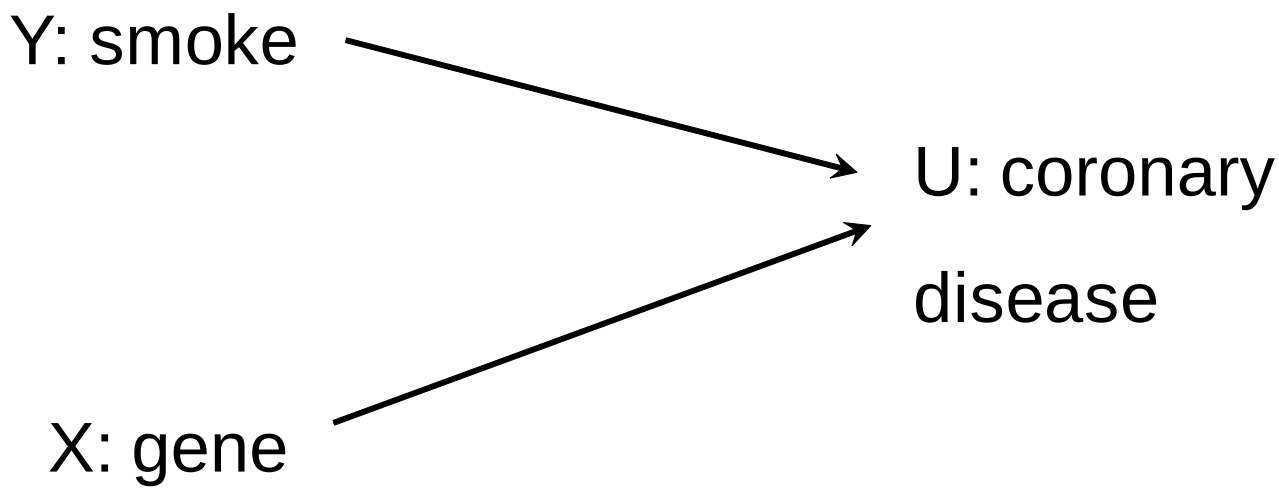
Z = {U} then path between X and Y blocked by Z

Z= { } then path between X and Y is unblocked by Z



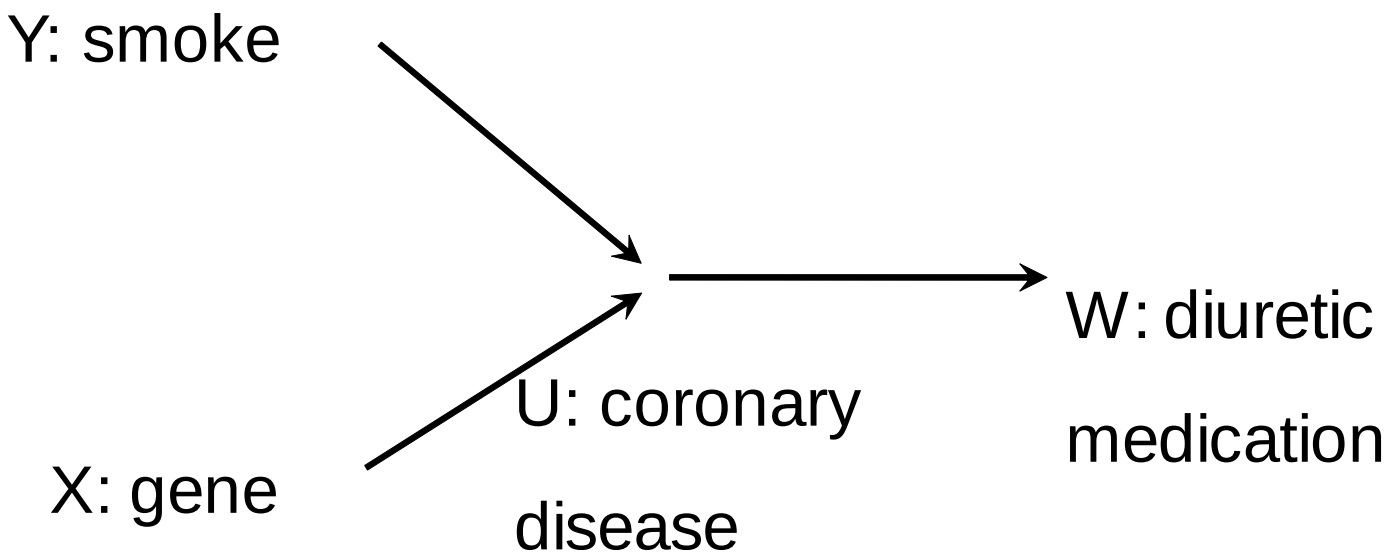
Z = {U} then path between X and Y blocked by Z

Z= { } then path between X and Y is unblocked by Z



Z = {U} then path between X and Y is **unblocked** by Z

Z= { } then path between X and Y is **blocked** by Z



Z = {W} then path between X and Y is **unblocked** by Z

Z = {U,W} then path between X and Y is **unblocked** by Z

Z= { } then path between X and Y is **blocked** by Z

Section II: Causal Diagrams and Structural Equation Models

- Structural equations models (SEM)
- Causal diagrams and causal DAG's
- Intervention DAG's and SEM's
- Counterfactuals
- Disturbance independence and the no-common causes assumptions

Structural equations

- Suppose that given $V = \{V_1, ..., V_k\}$,
  1. Each V_j is **determined by**:
     1. a **known subset** PA_j of $V - \{V_j\}$ and,
     2. **other** variables U_j .

- Denote the **deterministic** map between (PA_j, U_j) and V_j by

$$V_j = f_j(PA_j, U_j) \tag{3}$$

- (3) is called a **structural equation.**

- The variables U_j are called **disturbances or errors**

What makes an equation structural?

- Consider the following structural equations for T and S where
  - $S =$ indicator that the fasten your sit belt sign is on,
  - $T =$ the airplane experiences turbulences.

$$\begin{aligned} T &= U_T \\ S &= 1 - (1 - T)(1 - U_S) \end{aligned}$$

  - U_T is the indicator that a condition that generates a turbulence happened
  - U_S is the indicator that an event, other than turbulence, that prones the captain to turn on the sign, happened
- The system is algebraically equivalent to the system

$$\begin{aligned} S &= U_S^* \\ T &= S + U_T^* \end{aligned}$$

  with $U_S^* = 1 - (1 - U_T)(1 - U_S)$ and $U_T^* = -U_S(1 - U_T)$
- However, the equations in the first system are **structural** and the second are not? Why????

What makes an equation structural?

- The reason is because structural equations indicate the mechanisms by which the variables are created by nature. If the right hand side of the equation is a non-trivial function of a variable, then it means that nature will use that variable to create the variable in the left hand side of the equation.

- The equations

$$\begin{aligned} T &= U_T \\ S &= 1 - (1 - T)(1 - U_S) \end{aligned}$$

are structural because they tell us how nature "creates" T *from* S *and other factors* and how it creates S *from* T *and other factors.*

1. The first equation tells us that to "create" a turbulence, nature does not care if the sit belt sign is on.
2. The second equation tells us that to "make" a sit belt sign to be "on" it matters if there is a turbulence

What makes an equation structural?

- In contrast, the equations

$$\begin{aligned} S &= U_S^* \\ T &= S + U_T^* \end{aligned}$$

are not structural because

1. the first equation tells that the presence of an "ON" sign is not affected by the occurrence of a turbulence.
2. the second equation implies that the occurrence of a turbulence depends on whether or not the sign is on. In particular, the equation implies the ridiculous mechanism whereby a turbulence will always be formed when the sign is on.and the "external factor" U_T^* is 0.

Structural equations model

- **Definition:** A **structural equation model** (SEM) is a the model that assumes:

  1. a complete set of k **structural equations**

$$V_j = f_j\left(PA_j, U_j\right), \ j = 1, ..., k \tag{4}$$

     such that for each fixed value of $\left(U_1, ..., U_k\right)$, the system has a unique solution $V_1, ..., V_k$
  2. no element of $\left\{V_1, ..., V_k\right\}$ is a determinant of U_j for any j
  3. possibly, some facts about the determinants of the $U_j's$

- Examples of item 3
  1. no pair $\left(U_j, U_l\right)$ shares common determinants
  2. the pair $\left(U_j, U_l\right)$ only shares (unknown) common determinants
  3. U_j is a determinant of U_l
  4. U_j is equal to U_l

Types of structural equation models

- A SEM is further subclassified depending on the assumptions made about the $f_j's$

  1. If all $f_j's$ are **assumed to be unknown** then the model is called a **non-parametric structural equation model.**

  2. If all $f_j's$ are **assumed to be linear functions** of the $PA_j's$ and additive on the $U_j's$ then the model is called a **linear structural equation model.**

- The only assumptions encoded in a non-parametric SEM are the assumptions that the subset $V - PA_j$ does not participate in the construction of the variable V_j.

Causal diagrams

- **Definition**: Given a structural equation with variables $V_1, ..., V_k$, a *causal diagram* is a graph with nodes $V_1, ..., V_k$ such that it has

  1. a solid-line arrow from each node in the set PA_j to the node V_j, .for each j, and
  2. a dashed-line bidirected edge between any pair of nodes (V_j, V_k) unless the SEM assumes that
     1. the corresponding disturbances (U_j, U_k) do not share common determinants, and
     2. U_j is not a determinant of U_k
     3. U_k is not a determinant of U_j

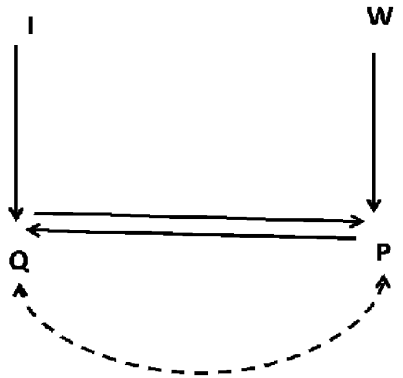
Remarks about causal diagrams

1. Causal diagrams are generally taken as a representation of the **associated non-parametric SEM**.

2. A causal diagram without double-dashed arcs is one in which every variable that is a common determinant of two other variables is included as a V variable of the system

Causal diagrams

- **Example 1: price and demand**



1. **Structural equations**

$$
\begin{aligned}
I &= f_I(U_I), && I = \text{ household income} \\
W &= f_W(U_W), && W = \text{ wage rate for producing product } A \\
Q &= f_Q(P, I, U_Q), && Q = \text{household demand for product } A \\
P &= f_P(Q, W, U_P), && P = \text{ unit price for product } A
\end{aligned}
$$

2. **Disturbance assumptions**. Only U_P and U_Q share common determinants

SEMs and Causal Diagrams

- Geneticist Sewall Wright (1921, 1934) was the first to use a system of (linear) equations combined with diagrams to communicate causal relationships.
- He was aware that equations alone were not satisfactory for encoding causal influences because any one equation implies other equations for the variables in the RHS which do not reflect the mechanism by which the variables are determined.
- Thus, his bright idea was to **append** to the equations the **causal diagram** which now reflected univocally the direction in which each equation ought to be read.

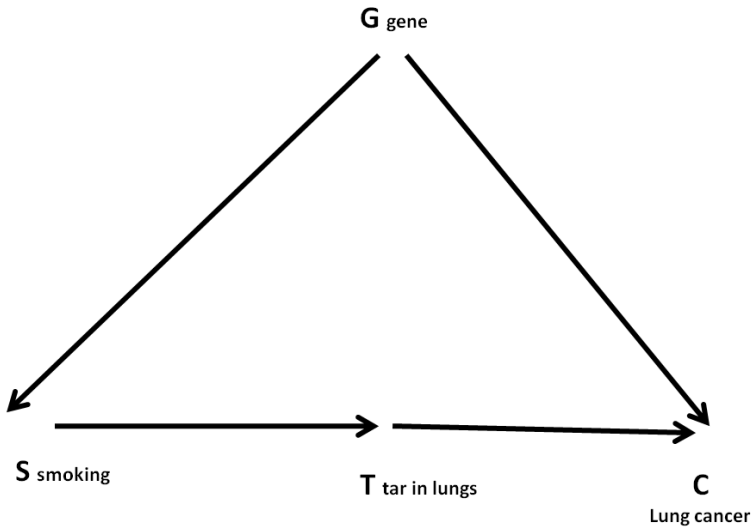
Recursive SEMs

- **Definition:** A **recursive SEM or Semi-Markovian SEM** is a SEM whose causal diagram is such that when its double-dashed arrows are deleted, the resulting graph is a **DAG.**

- **Property 1:** In a recursive SEM: $V_l \in PA_j \Rightarrow V_j \notin PA_l$

- **Property 2**: In a recursive SEM there exists an ordering $V_1, ..., V_k$ such that given $U = \{U_1, ..., U_k\}$, the variables in V are determined recursively, V_1 first, V_2 next, and so on.

- **Example 1: smoking and lung cancer**

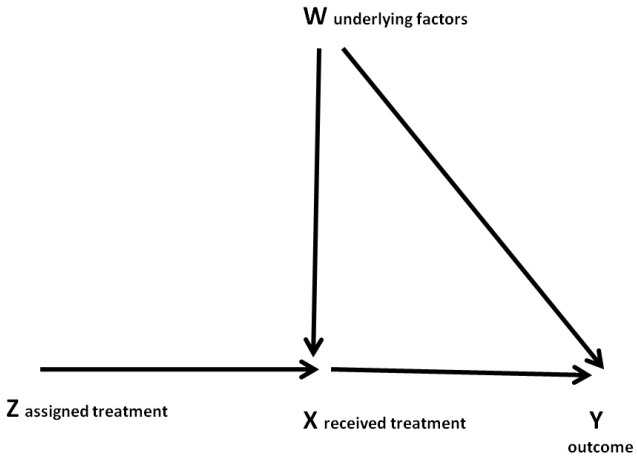


1. **Structural equations**

$$
\begin{aligned}
G &= f_G(U_G), && G = \text{ genetic trait} \\
S &= f_S(G, U_S), && S = \text{ smoking indicator} \\
T &= f_T(S, U_T), && T = \text{amount of tar accumulated in the lung} \\
C &= f_C(G, T, U_C), && C = \text{ indicator of lung cancer}
\end{aligned}
$$

2. **Disturbance assumptions**. No pair of disturbances share a common determinant

- **Example 2: non-compliance in clinical trials**

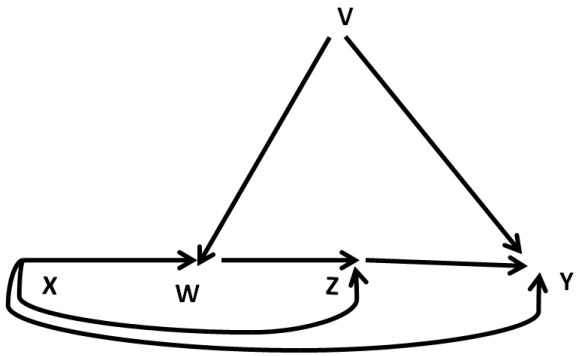


1. **Structural equations**

$$
\begin{aligned}
W &= f_W(U_W), \quad W = \text{factors affecting compliance and response} \\
Z &= f_Z(U_Z), \quad Z = \text{treatment assigned} \\
X &= f_X(Z, W, U_X), \quad X = \text{treatment received} \\
Y &= f_Y(X, W, U_Y), \quad Y = \text{health outcome}
\end{aligned}
$$

2. **Disturbance assumptions**. No pair of disturbances share a common determinant. Note that Z is not determined by any other variable because treatment assignment has been randomized.

- **Example 3: sequentially randomized clinical trial.** Full randomization of treatment X and randomization to Z with probability that depends on observed health history and first assigned treatment



- **SEM: jointly independent disturbances and**

$$
\begin{aligned}
V &= f_V(U_V), && V = \text{ immune status} \\
X &= f_X(U_X), && X = \text{ treatment randomized at baseline} \\
W &= f_W(X, V, U_W), && W = \text{response after first treatment} \\
Z &= f_Z(X, W, U_Z), && Z = \text{ second randomized treatment} \\
Y &= f_Y(Z, X, V, U_Y) && Y = \text{response at end of study}
\end{aligned}
$$

SEM $\longrightarrow$ CAUSAL DIAGRAM

RECURSIVE SEM $\longrightarrow$ CAUSAL DIAGRAM IS DAG + DASHED DOUBLE ARROWS

RECURSIVE SEM + NO COMMON CAUSES FOR THE ERRORS $\longrightarrow$ **CAUSAL DIAGRAM IS DAG**

Probabilistic SEM

- A **probabilistic structural equation model** is a SEM in which the disturbances $U = (U_1, ..., U_k)$ are assumed to be **random variables**.
- Of course, if $U_j, j = 1, ..., k$, is a random variable, then so are the variables $V_j, j = 1, ..., k$, of the SEM.
- The distribution $p(u)$ of U and a fixed set of structural functions $f_j, j = 1, ...k$, uniquely determine the distribution of $p(v)$ of $V = (V_1, ..., V_k)$.
- If U is generated by nature with distribution $p(u)$, then V is generated by nature with law $p(v)$.
- $p(v)$ is called the **observational law** of V

Intervention SEM

- A key implicit assumption of SEMs is that **modification** of one equation **alters** the values of the inputs to other equations but **not the functional form** of the equations themselves

- In a SEM each equation represents an **isolated mechanism,** if you **intervene and modify** one mechanism you do not change the others

Intervention SEM

- A recursive SEM is like an electrical circuit with black boxes, the j^{th} one receiving the input (PA_j, U_j) and spitting the output V_j.

- If you were to **intervene** and replace **one specific** black box with another one, your **action** will have the effect of altering the input of the **boxes connected to the replaced box** but your action will **not affect** (i.e. alter) **any of these boxes.**

Intervention SEM

- This means that if you intervene to modify the mechanism that creates one variable, you will modify
  - neither the *equations* (i.e. mechanisms) that dictate the creation of the remaining variables in the system nor,
  - the values of the disturbances (as they are determined by factors outside the system).
- So we can define a new SEM representing how the variables V would be created in the hypothetical world in which we intervene and force a subset of V to be fixed at given values.
- In such SEM we simply replace the equations that create the intervened variables with new equations in which each variable is equal to the given constant

Intervention SEMs

- **Definition:** given a SEM

$$V_j = f_j\left(PA_j, U_j\right), j = 1, ..., k$$

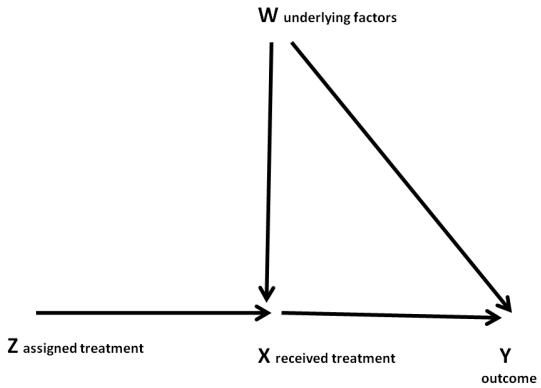
an **intervened SEM with intervened variables** V_{j_l} set to v_{j_l} , $l = 1, ..., l^*$ is a **new SEM** defined by the structural equations

$$\begin{aligned} V_j &= f_j\left(PA_j, U_j\right), j \notin \left\{j_1, ..., j_{l^*}\right\} \\ V_{j_l} &= v_{j_l}, l = 1, ..., l^* \end{aligned}$$

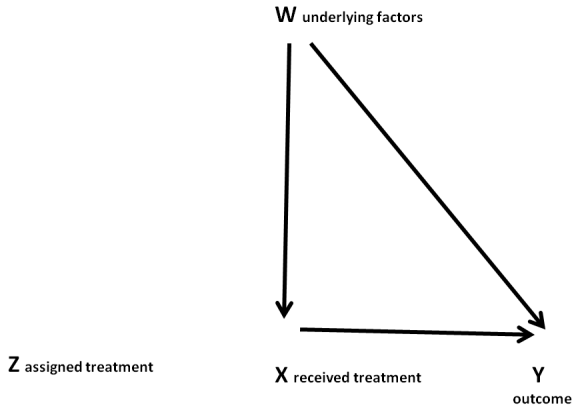
- The **causal diagram of an intervention SEM** is identical to one of the original SEM but in which all arrows pointing to the intervened variables (including any dashed double-edges pointing to it, if they exist) are removed.

Intervention causal diagrams

- **Example:** suppose that we intervene in the system represented by the DAG



to force $X = x$. Then the intervened DAG is



Counterfactual variables and intervention distributions

- Consider a probabilistic intervened SEM in which we **intervene to set** X to x.
- We denote the variables solving the new system with

$$V_x = (V_{x,1}, ..., V_{x,k})$$

- The variables $V_{x,j}$ are referred to as **potential variables or counterfactuals.**
- We define the **intervention distribution**

$$\boxed{p_x(v) \equiv \Pr(V_x = v)}$$

Counterfactual variables and intervention distributions

- Note that the **intervention distribution**

$$\boxed{p_x(v) \equiv \Pr(V_x = v)}$$

is the probability that we would observe that the left hand side variables of SEM be equal to v in a world in which we **impose** the action $X = x$ on every possible realization of the disturbances U.

- This law is **NOT** generally equal to

$$\boxed{p(v|x) \equiv \Pr(V = v|X = x)}$$

which is the **conditional probability** that $V = v$ given $X = x$. This

is the probability that $V = v$ among those that **we observe** to have $X = x$

Condl vs intervention distbs are not the same. Example.

- Consider the SEM

$$\boxed{Z = U_z, \quad X = Z + U_x}$$

$U_Z \amalg U_X$ both Bernoulli with success probabilities π_z and π_x.

- Then, for $v = (z, x) = (1, 1)$, we have

$$\begin{aligned} \Pr(V = v | X = x) &= \frac{\Pr(Z = 1, X = 1)}{\Pr(X = 1, Z = 1) + \Pr(X = 1, Z = 0)} \\ &= \frac{\pi_z (1 - \pi_x)}{\pi_z (1 - \pi_x) + \pi_x (1 - \pi_z)} \end{aligned}$$

- On the other hand, $p_x(v) = \Pr(Z_x = 1)$ is the probability that $Z = 1$ under the modified SEM

$$\boxed{Z = U_z, \quad X = 1}$$

But in this system, $Z = 1$ with probability π_z, so $p_x(v) = \pi_z$.

Independence and the no-common causes assumption

- **Assumption**: if the causal diagram of a recursive probabilistic SEM has **no dashed bi-directed edges** then the disturbances $U_1, ..., U_k$ are **mutually independent**..

- Recall that a causal diagram without double-dashed arcs is one in which every variable that is a common determinant of two other variables is included as a V variable of the system

Markovian SEMs

- **Definition:** a **Markovian SEM i**s a probabilistic recursive SEM whose causal diagram does not have **dashed bi-directed edges, i.e. it is a DAG**.

- **Property:** if a SEM is Markovian, then any intervention SEM derived from it is also Markovian.

- **Proof:** immediate. The error vector U is the same in the original and the intervention SEM of a recursive SEM is also recursive.

Section III: identifiability of the intervention law, preliminaries

- The Causal Markov Condition
- The positivity condition
- Trimmed graphs
- The three rules of the "do calculus"
    - The back-door theorem

Causal Markov condition

- **Theorem (the causal Markov condition)**: The DAG of a Markovian SEM

$$V_j = f_j\left(PA_j, U_j\right), j = 1, ..., k$$

represents the joint law of the variables $V = V_1, ..., V_k$, i.e.

$$p\left(v\right) = \left\{\prod_{j=1}^{k} p\left(v_j | pa_j\right)\right\} I_{\{p(\cdot)>0\}}\left(v\right)$$

Proof of the causal Markov condition

- **Proof:** Let the order $V_1, ..., V_k$ be consistent with the DAG. Then, independence of the errors and recursiveness implies that

$$U_j \amalg \overline{V}_{j-1} \tag{5}$$

where $\overline{V}_{j-1} = (V_1, ..., V_{j-1})$. Then,

$$p(v) = \left\{ \Pi_{j=1}^{k} \Pr\left(V_j = v_j | \overline{V}_{j-1} = \overline{v}_{j-1}\right) \right\} I_{\{p(\cdot)>0\}}(v)$$

But

$$\begin{aligned} \Pr\left(V_j = v_j | \overline{V}_{j-1} = \overline{v}_{j-1}\right) &= \Pr\left(f_j(pa_j, U_j) = v_j | \overline{V}_{j-1} = \overline{v}_{j-1}\right) \\ &= \Pr\left(f_j(pa_j, U_j) = v_j\right) \text{ by } (5) \\ &= g_j(v_j, pa_j) \end{aligned}$$

Which proves that $\Pr\left(V_j = v_j | \overline{V}_{j-1} = \overline{v}_{j-1}\right)$ depends only on pa_j, hence

$$\Pr\left(V_j = v_j | \overline{V}_{j-1} = \overline{v}_{j-1}\right) = \Pr\left(V_j = v_j | PA_j = pa_j\right)$$

This concludes the proof...

The positivity condition

- Our next Theorem establishes that if the following positivity condition holds, $p_{x'}(\cdot)$ is identified (i.e. it is a functional of) the observational law $p(\cdot)$ of V.

- **The positivity condition for** $X = x'$. Given a Markovian SEM with variables V, a subset $X = \{X_1, ...X_l\}$ of V, and a fixed constant vector $x' = (x'_1, ..., x'_l)$, it holds that for every pa_j such that $\Pr(PA_{X_j} = pa_j) > 0$,

$$\Pr\left(X_j = x'_j | PA_{X_j} = pa_j\right) > 0, j = 1, ..., l \tag{6}$$

- The condition stipulates that, regardless of the values of the parents of X_j, in the observational world there is always a positive chance that X_j will take the selected value x'_j.

The identification theorem

- **Theorem (identification):** if the positivity condition for $X = x'$ holds then $p_{x'}(\cdot)$ is absolutely continuous with respect to $p(\cdot)$ .and

$$\boxed{p_{x'}(v) = \left\{ \Pi_{j:v_j \notin x'} p(v_j | pa_j) \times I_{\{x'\}}(x) \right\} I_{\{p(\cdot)>0\}}(v)} \qquad (7)$$

Equivalently, the likelihood ratio satisfies

$$\boxed{\frac{p_{x'}(v)}{p(v)} I_{\{p(\cdot)>0\}}(v) = \frac{I_{\{x'\}}(x)}{\Pi_{i=1}^{s} p(x_i | pa_i)} I_{\{p(\cdot)>0\}}(v)}$$

- **Definition**: The formula on the right hand side of (7) is called **the intervention formula.**

Remarks on the identification theorem

- The intervention formula

$$\left\{\Pi_{j:v_j \notin x'} p\left(v_j | pa_j\right) \times I_{\{x'\}}\left(x\right)\right\} I_{\{p(\cdot)>0\}}\left(v\right)$$

  is a functional of $p\left(\cdot\right)$

- **Corollary**: if the positivity condition for $X = x'$ holds,

> If all the variables V are measured $\Rightarrow$
> $p\left(\cdot\right)$ can be estimated consistently $\Rightarrow$
> $p_{x'}\left(\cdot\right)$ can be estimated consistently

Identifiability from a subset of the nodes of a causal DAG.

- In practice, however, only a subset B of the variables in the causal DAG are measured and we can only hope to estimate consistently $p(b)$.
- Hence we can estimate consistently $p_x(y)$ if it depends on $p(v)$ only through $p(b)$ but not otherwise
- The following question is then ultra important in practice.

> Suppose that in a causal DAG, $B \subset V, X \subset B$, $Y \subset B$ and $X \cap Y = \varnothing$
> What are sufficient conditions under which the intervention law $p_x(y)$ is a functional of $p(b)$ only?

.

Sufficient conditions for identification

- There exist a number of graphical rules that one can use to check for such sufficient conditions for identifiability.
- The sufficient conditions are derived from three key graphical results for causal DAGs, known as the *rules of the do (or intervention) calculus*. So we will start by stating these rules
- The rules are indeed Theorems and they are proved in Pearl (*1995*, *Biometrika*).

Trimmed graphs: preliminary notation.

- Let X, Y and Z be arbitrary disjoint sets of nodes of a DAG G.
- **Convention 1:** $G_{\overline{X}}$ is the graph obtained by deleting from G all arrows pointing to nodes in X
- **Convention 2:** $G_{\underline{X}}$ is the graph obtained by deleting from G all arrows emerging from nodes in X
- **Convention 3:** $G_{\overline{X},\underline{Z}}$ is the graph obtained by deleting from G all arrows pointing to nodes in X and all arrows emerging from nodes in Z

Rules of do calculus (Adapted from Pearl, Biometrika, 1995)

- Let Y, Z and W be disjoint subsets in a causal DAG G.
- **Rule 1: d-separation.** *(not really a causal result)*

$$\boxed{\text{if } (Y \amalg Z|W)_G \text{ then } p(y|z,w) = p(y|w),}$$

- **Rule 2: back-door** *(when is observing the same as intervening).* Suppose

$$\boxed{\text{if } (Y \amalg Z|W)_{G_{\underline{Z}}} \text{ then } p_z(y|w) = p(y|z,w)}$$

for all (z,w) such that $p(z,w) > 0$

- **Rule 3: action irrelevance** *(about actions that have no effects)*

$$\boxed{\text{if } (Y \amalg Z)_{G_{\overline{Z}}} \text{ then } p_z(y) = p(y)}$$

Rules of do calculus in terms of counterfactuals

- **Rule 1:d-separation.** *(not really a causal result)*

$$\text{if } (Y \amalg Z|W)_G \text{ then.}$$
$$\Pr(Y = y|Z = z, W = w) = \Pr(Y = y|W = w)$$

- **Rule 2: back-door** *(when is observing the same as intervening)*

$$\text{if } (Y \amalg Z|W)_{G_{\underline{Z}}} \text{ then}$$
$$\Pr(Y_z = y|W_z = w) = \Pr(Y = y|Z = z, W = w)$$

- **Rule 3: action irrelevance** *(about actions that have no effects)*

$$\text{if } (Y \amalg Z)_{G_{\overline{Z}}} \text{ then}$$
$$\Pr(Y_z = y) = p(Y = y)$$

Remark about the rules as Pearl stated them

- Pearl stated the rules not quite as we did.
- Rule 3 in Pearl (1995) is slightly more general. Also,
- Pearl used
  1. $G_{\underline{X}}$ instead of G
  2. $p_{x,z}$ instead of p_z, and
  3. p_x instead of p
- His results are just a re-statement of ours when we regard the "observational" DAG as the DAG with X intervened at x and the observational p as the intervention law p_x

Let's recall the rules

- **Rule 1:d-separation.***(not really a causal result)*

$$\text{if } (Y \amalg Z|W)_G \text{ then.}$$
$$\Pr(Y = y|Z = z, W = w) = \Pr(Y = y|W = w)$$

- **Rule 2: back-door** *(when is observing the same as intervening)*

$$\text{if } (Y \amalg Z|W)_{G_{\underline{Z}}} \text{ then}$$
$$\Pr(Y_z = y|W_z = w) = \Pr(Y = y|Z = z, W = w)$$

- **Rule 3: action irrelevance** *(about actions that have no effects)*

$$\text{if } (Y \amalg Z)_{G_{\overline{Z}}} \text{ then}$$
$$\Pr(Y_z = y) = p(Y = y)$$

Rule 2

$$\text{If } \left(Y \amalg Z | W\right)_{G_{\underline{Z}}} \text{ then}$$

$$p_z\left(y|w\right) = p\left(y|z,w\right)$$

or equivalently

$$\Pr\left(Y_z = y | W_z = w\right) = \Pr\left(Y = y | Z = z, W = w\right)$$

- In $G_{\underline{Z}}$ the only paths from Z to Y are through paths that start with an edge that points into Z. These paths are called **back-door paths**.
- The condition $\left(Y \amalg Z | W\right)_{G_{\underline{Z}}}$ says that **all back-door paths from Z to Y are blocked by W**.
- The essential part of **Rule 2** is so important, that it deserves the qualification of *Theorem*. We re-state it as such now.

The back-door theorem

- **Theorem:** Let Y, Z and W be three disjoint set of nodes in a causal DAG Γ. Then for all $(z, w) : p(z, w) > 0$,

$$p_z(y|w) = p(y|z, w)$$

or equivalently

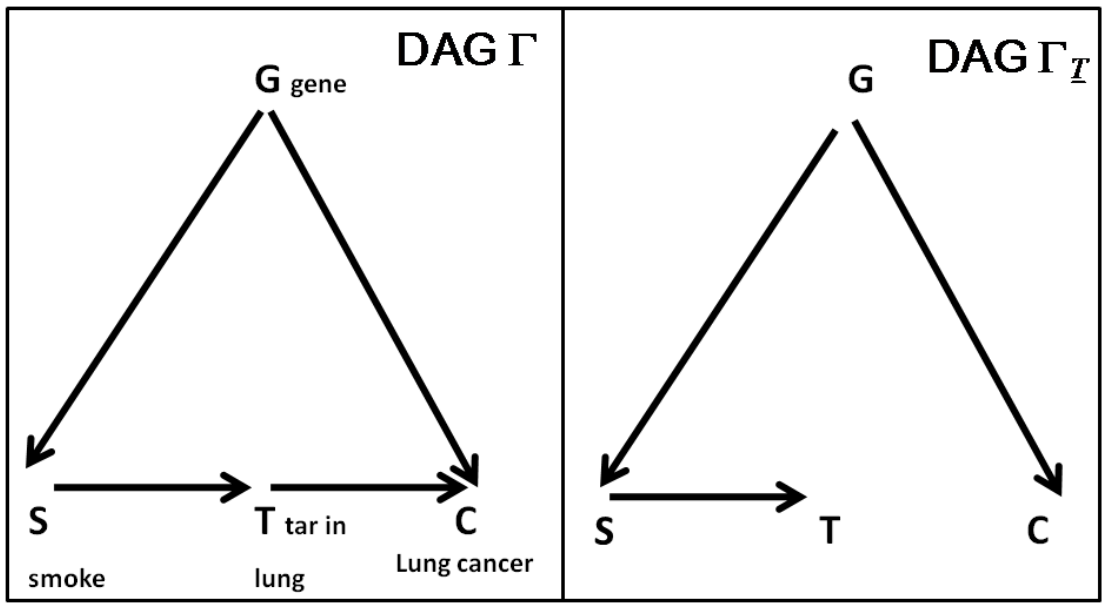
$$\Pr(Y_z = y|W_z = w) = \Pr(Y = y|Z = z, W = w)$$

if all back-door paths from Z to Y are blocked by W.

Example of Rule 2

Back door path between T and C is T, S, G, C which is blocked by $G \Rightarrow$

$$\Pr\left(C_t = c \middle| G_t = g\right) = \Pr\left(C = c \middle| T = t, G = g\right)$$



Let's recall the rules

- **Rule 1:d-separation.** *(not really a causal result)*

$$\text{if } (Y \amalg Z|W)_G \text{ then.}$$
$$\Pr(Y = y|Z = z, W = w) = \Pr(Y = y|W = w)$$

- **Rule 2: back-door** *(when is observing the same as intervening)*

$$\text{if } (Y \amalg Z|W)_{G_{\underline{Z}}} \text{ then}$$
$$\Pr(Y_z = y|W_z = w) = \Pr(Y = y|Z = z, W = w)$$

- **Rule 3: action irrelevance** *(about actions that have no effects)*

$$\text{if } (Y \amalg Z)_{G_{\overline{Z}}} \text{ then}$$
$$\Pr(Y_z = y) = p(Y = y)$$

Remark on Rule 3

- In the DAG $G_{\overline{Z}}$ the only unblocked paths between Z and Y are the directed paths paths between Z and Y in G
- The condition $(Y \amalg Z)_{G_{\overline{Z}}}$ is then the condition that in DAG G there are no directed paths between Y and Z
- The conclusion $\Pr(Y_z = y) = p(Y = y)$ implies that Z has no causal effect on Y *(if we intervene to set $Z = z$, then regardless of the value z at which we set Z, the distribution of the outcome will be the same)*
- Then the result

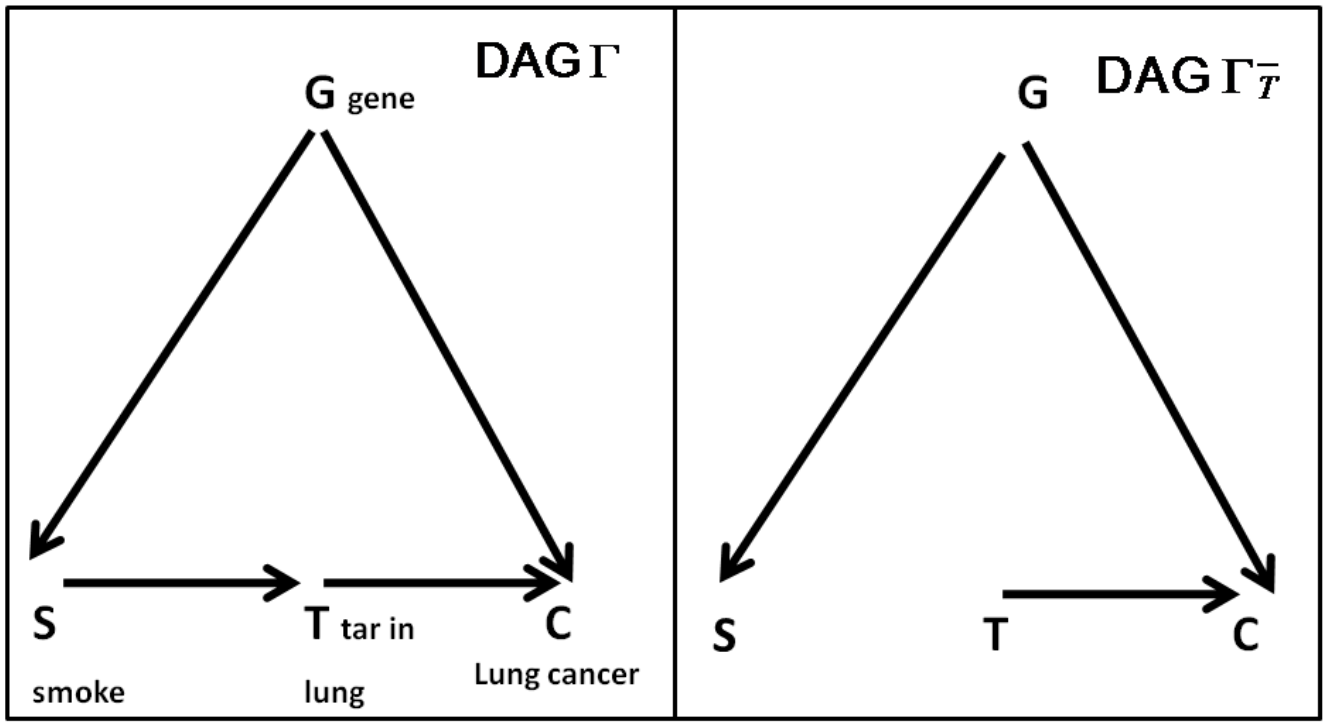
$$\boxed{\text{if } (Y \amalg Z)_{G_{\overline{Z}}} \text{ then } \Pr(Y_z = y) = p(Y = y)}$$

implies that if in the original DAG there is no directed path between Z and Y then Z *has no causal effect on* Y.

First example of Rule 3.

- **Future actions don't affect past outcomes (***reducing the tar in your lungs will not reduce how much you smoke)*

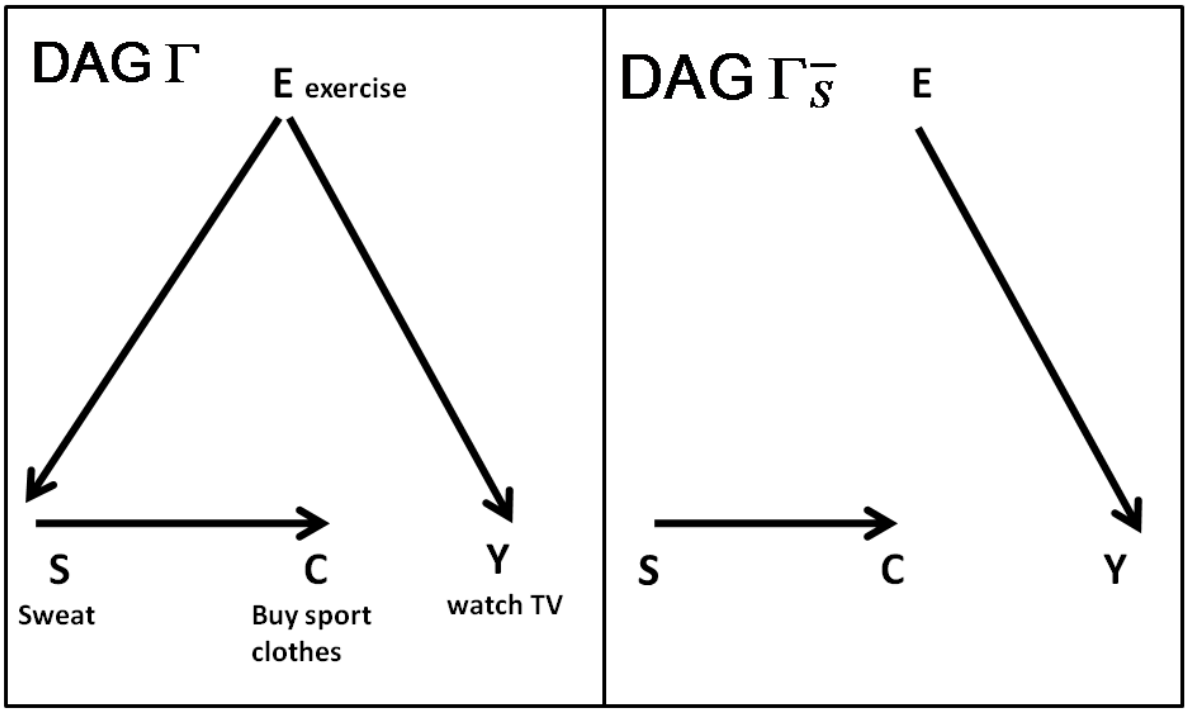
$$(S \amalg T)_{\Gamma_{\overline{T}}} \Rightarrow \Pr(S_t = s) = \Pr(S = s)$$



Second example of Rule 3.

- **Actions without effects** *(your sweating does not cause your inclination-or not- to watch TV)*

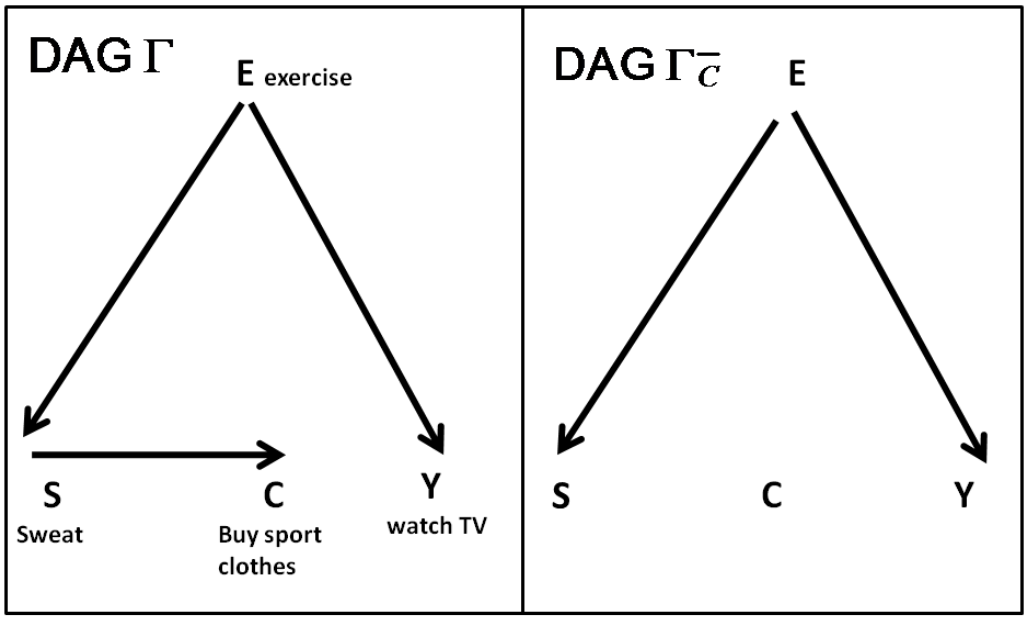
$$(S \amalg Y)_{\Gamma_{\overline{S}}} \Rightarrow \Pr(Y_s = y) = \Pr(Y = y)$$



Second example of Rule 3.

- **Actions without effects** *(your inclination - or not- to buy sport clothes does not cause your inclination -or not- to watch TV)*

$$(C \amalg Y)_{\Gamma_{\overline{C}}} \Rightarrow \Pr(Y_c = y) = \Pr(Y = y)$$



Section IV: identifiability of the intervention law: the back-door theorem

- The back-door adjustment theorem
  - the intervention formula
  - standardized vs crude rates
  - the regression and the inverse probability weighted forms
  - the propensity score
- Lessons from the back-door theorem
  - measuring all common causes of treatment and outcome is not always needed
  - it is not always ok to adjust for proxies of common causes of treatment and outcome
  - it is not always ok to adjust for common correlates of treatment and outcome
  - Berkson bias
    - M-structures
    - Drop-out in longitudinal studies

Corollaries of the "do" calculus: the back-door adjustment

- **Theorem (the back-door adjustment):** let X, Y and Z be disjoint set of nodes in a causal DAG G and suppose that (x, z) are fixed values such that $p(x, z) > 0$. If Z is a non-descendant of X that blocks all back doors between X and Y then

$$\boxed{p_x(y, z) = p(y|x, z)\, p(z)}$$

- **Proof:** for (x, z) such that $p(x, z) > 0$ we have

$$\begin{aligned} & p(y|x, z)\, p(z) = \\ & = p_x(y|z)\, p(z) \;\; \textit{by the back-door theorem} \\ & = p_x(y|z)\, p_x(z) \;\; \textit{by rule 3 } (Z \textit{ is non-descendant of } X) \\ & = p_x(y, x) \end{aligned}$$

Corollaries of the "do" calculus: the back-door adjustment

- **Corollary 1:** under the assumptions of the theorem

$$\boxed{p_{x'}(y,z,x) = p(y|x,z)\,p(z)\,I_{\{x'\}}(x)}$$

or equivalently

$$\boxed{\frac{p_{x'}(y,z,x)}{p(y,x,z)} I_{\{p()>0\}}(y,x,z) = \frac{I_{\{x'\}}(x)}{p(x|z)} I_{\{p()>0\}}(y,x,z)}$$

So we reproduce the *intervention formula* for the subset $Y \cup X \cup Z$ of the variables in the DAG!!

Corollaries of the "do" calculus: the back-door adjustment

- **Corollary: Under the conditions of the theorem,**

$$\boxed{p_x(y) = \sum_z p(y|x,z)\, p(z)}$$

Which variables we need to identify treatment effects?

- It follows from the preceding theorem that to identify $p_x(y)$ we don't need to measure all variables in a causal DAG.
- It suffices to measure, besides Y and X, a set Z that
  1. are non-descendants of X and,
  2. block all the back doors between X and Y.
- Variables Z that satisfy the two preceding conditions are said to satisfy *the back-door criterion*

Standardized vs crude risks

- The back-door theorem says that if Z satisfies the back-door criterion then

$$\Pr(Y_x = y) = \underbrace{\sum_z \underbrace{\Pr(Y = y|X = x, Z = z)}_{\text{crude stratum-specific rates}} \times \underbrace{\Pr(Z = z)}_{\text{weights}}}_{\text{standardized rate}}$$

standardized rate: weighted averaged of stratum specific crude rates
weights are strata prob. in the population

- This is different from

$$\Pr(Y = y|X = x) = \underbrace{\sum_z \Pr(Y = y|X = x, Z = z) \underbrace{\Pr(Z = z|X = x)}_{\text{weights}}}_{\text{crude rate}}$$

crude rate: weighted averaged of stratum specific crude rates
weights are strata prob. in the supopul. with X equal x

The regression and the IPW forms

- We have seen that when Z meets the back-door criterion

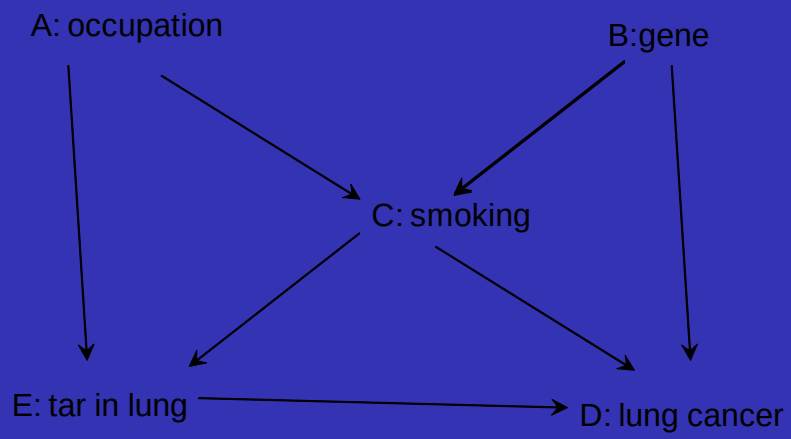
$$p_x(y) = \sum_z p(y|x,z)\,p(z) \text{ and}$$

$$\frac{p_{x'}(y,z,x)}{p(y,x,z)} = \frac{I_{\{x'\}}(x)}{p(x|z)}$$

- This implies that

$$\boxed{\begin{aligned} E(Y_x) &= E\{E(Y|X=x,Z)\} \\ &= E\left\{\frac{I_{\{x\}}(X)}{\Pr(X=x|Z)}Y\right\} \end{aligned}}$$

- The expressions in the RHS are **two forms of the SAME functional** of $p(y,x,z)$
  1. The first expression is called **the regression form**
  2. The second expression is called **the inverse probability weighted form**
- $\pi(z) \equiv \Pr(X=x|Z=z)$ is called the **propensity score** for trx x



- We will next examine which variables satisfy the back door criterion for the pair (E, D)
  1. A does not satisfy it because it does not block the path E, C, D
  2. B does not satisfy it for the same reason
  3. C does not satisfy it because it unblocks the path E, A, C, B, D
  4. (A, C) satisfies it!!. Also, (B, C) satisfies it!!

First lesson: measuring all common causes is not always needed.

- Thus, we conclude that

$$\begin{aligned} p_e(d) &= \sum_a \sum_c p(d|e,a,c)\, p(a,c) \\ &= \sum_b \sum_c p(d|e,b,c)\, p(b,c) \end{aligned}$$

- Thus, to identify $p_e(d)$ it suffices to measure
  - the variables A, C, E, D or
  - the variables B, C, E, D.
- But we don't need to measure all three common causes A, B and C !!!!
- This exemplifies how DAGs can be used to help design studies!

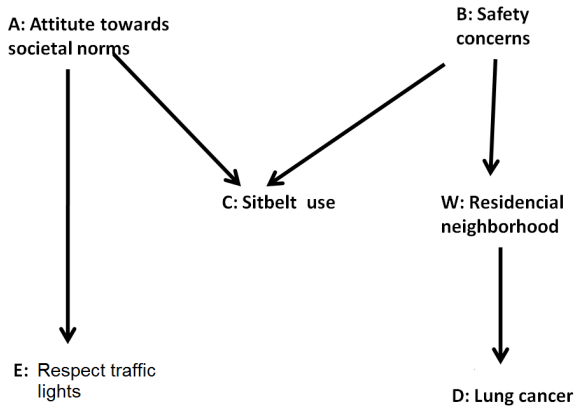
Second lesson: it is not OK to adjust for proxies of unmeasured common causes

- Measuring just A, E, D or just B, E, D or just C, E, D will not suffice to identify $p_e(d)$.
- In particular, in general,

$$p_e(d) \neq \sum_a p(d|e,c)\, p(c)$$

- C is a proxy for (i.e. is correlated with) A and B.
- This example shows that it is NOT always OK to adjust for *proxies* of *unmeasured common causes*

Third lesson: it is not always ok to adjust for common correlates of exposure and disease



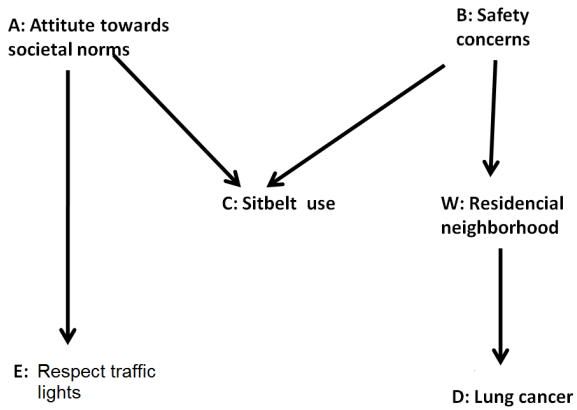
- C. is correlated with E and D, but

$$p_e(d) = p(d|e) \text{ by rule 2 because } (E \amalg D)_{G_{\underline{E}}} \Rightarrow$$
unadjusted rates are correct (no need to measure anything!)

- However, C unblocks the path E, A, C, B, W, D, thus in general,

$$p_e(d) \neq \sum p(d|e,c)\, p(c) \Rightarrow \textbf{adjustment for C is incorrect}$$

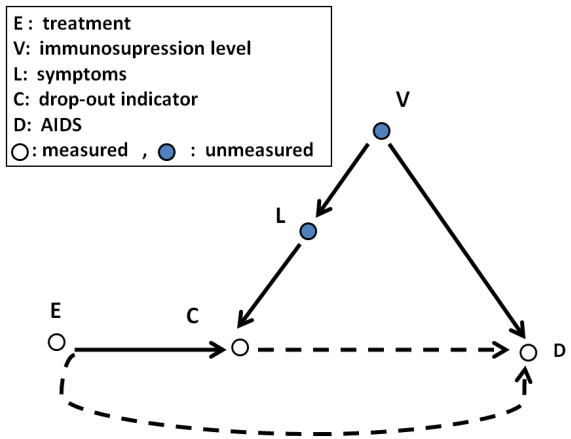
Fourth lesson: Berkson bias



- The structure of this DAG is known as an *M-structure.*
- The spurious correlation between D and E was induced because **we conditioned on a collider** (C)
- Any spurious correlation induced by *conditioning on colliders* is called **Berkson bias**

Other Berkson biases: drop-out in longitudinal studies

- Consider the following clinical trial of HIV+ patients



- We would like to compute

$$p_{e,c=0}\left(d\right)$$

the rate of disease in the hypothetical world in which everybody took $E = e$ and nobody dropped out

The "story" behind the previous DAG

- Patients are randomized to treatment or control (E) (E is a root node because of randomization)
- Patients in the treatment arm are at greater risk of side effects (nausea, vomiting, etc) and hence of dropping out (arrow from E to C)
- The greater the level of immunosuppression,
  1. the greater the risk of AIDS (arrow from U to D)
  2. the greater the risk of developing symptoms (fever, weight loss, etc) (arrow from U to D)
- The greater the risk of symptoms the greater the risk of dropping out (arrow from L to C)

Drop-out in longitudinal studies

- If in the true DAG the dashed arrows are absent, then there is no directed path from (E, C) to D so

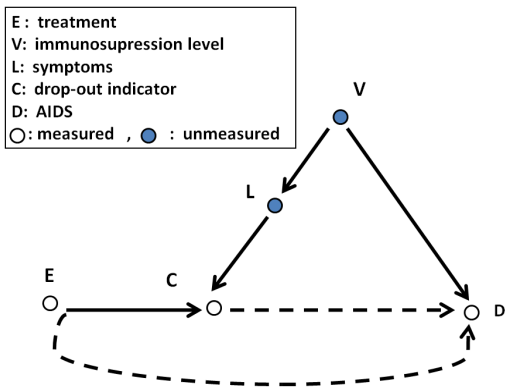
$$\boxed{p_{e,c=0}(d) = p(d) \text{ does not depend on } e}$$

- However, in general,

$$\boxed{p(d|e, c = 0) \text{ depends on } e}$$

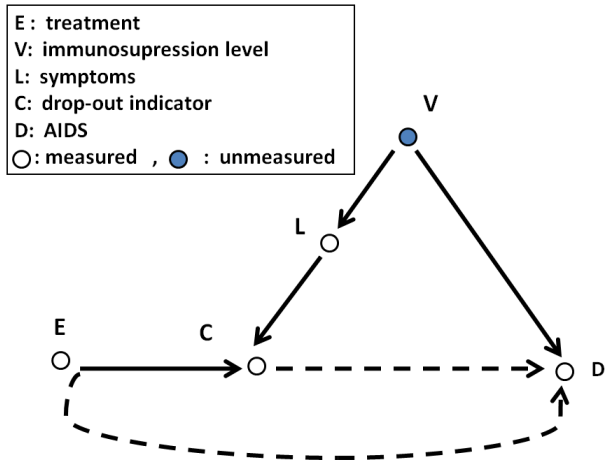
because the path E, C, L, V, D is unblocked by C

- **Conclusion**: restricting the analysis to patients for whom D *is not missing*, leads us to **incorrectly conclude** that E has an effect on D



Drop-out in longitudinal studies

- The effect of $(E, C = 0)$ is not identified if in the trial we only measure E, C and D
- However, if we also measure L



- Then L blocks all back-doors between (E, C) and D and we have that

$$\boxed{p_{e,c=0}(d) = \sum_l p(d|e, c = 0, l)\, p(l)}$$

Connections with the missing data literature

- In our example, the fact that E is a root node implies (by rule 3) that

$$p_{e,c=0}(d) = p_{c=0}(d|e)$$

- So, the mistake in using $p(d|e, c = 0)$ to estimate the effect of E on D is to assume that

$$p(d|e, c = 0) = p_{c=0}(d|e) \tag{8}$$

- In the missing data literature, (8) is known as the *assumption* -MCAR- that the D is missing completely at random conditional on E.

Connections with the missing data literature

- We now see that MCAR is tantamount to assuming that there are no common causes of missingness and disease, an often *very very unrealistic assumption*
- Notice that the problem of missing D is not resolved by imputing it from the law $p\left(d|e, c=0\right)$
- This imputation will only aggravate the problem because it will make you believe that (your biased) estimator is very precise thus giving you more confidence that your incorrect analysis is correct!
- Imputing garbage observations only helps improve the efficiency of estimators of garbage quantities!!!

Connections with the missing data literature

- The variable L does not intervene in the expression $p_{c=0}\left(d|e\right)$.
- However, to be able to identify $p_{c=0}\left(d|e\right)$ we need to have measured L because

$$\boxed{p_{c=0}\left(d|e\right)=\sum_{l}p\left(d|e,c=0,l\right)p\left(l\right)}$$

- In the missing data literature L is called an *auxiliary variable*, because it is a variable that does not intervene in the estimand of interest but that is needed to estimate it.

Connections with the missing data literature

- In our DAG L and E are $d-$separated, so $p\left(l|e\right)=p\left(l\right)$. Thus,

$$\boxed{p_{c=0}\left(d|e\right)=\sum_{l}p\left(d|e,c=0,l\right)p\left(l|e\right)} \tag{9}$$

- This is just the formula for the conditional probability of D given E under

$$\boxed{p_{c=0}\left(d,l,c'|e\right)=p\left(d|l,c',e\right)I_{\{0\}}\left(c'\right)p\left(l|e\right)}$$

- From where it follows that the likelihood ratio between the observed and the intervention laws (conditional on E) satisfies

$$\boxed{\frac{p_{c=0}(d,l,c'|e)}{p(d,l,c'|e)}=\frac{I_{\{0\}}(c')}{\Pr(C=0|E=e,L=l)}} \tag{10}$$

Connections with the missing data literature

- From

$$p_{c=0}\left(d|e\right)=\sum_{l}p\left(d|e,c=0,l\right)p\left(l|e\right) \tag{11}$$

we obtain

$$\underbrace{E_{c=0}\left(D|E=e\right)}_{\substack{\text{mean of } D \text{ given } E=e \\ \text{if nobody dropped out}}}=\underbrace{E\left\{E\left(D|E=e,C=0,L\right)|E=e\right\}}_{\text{the regression functional}}$$

- and from

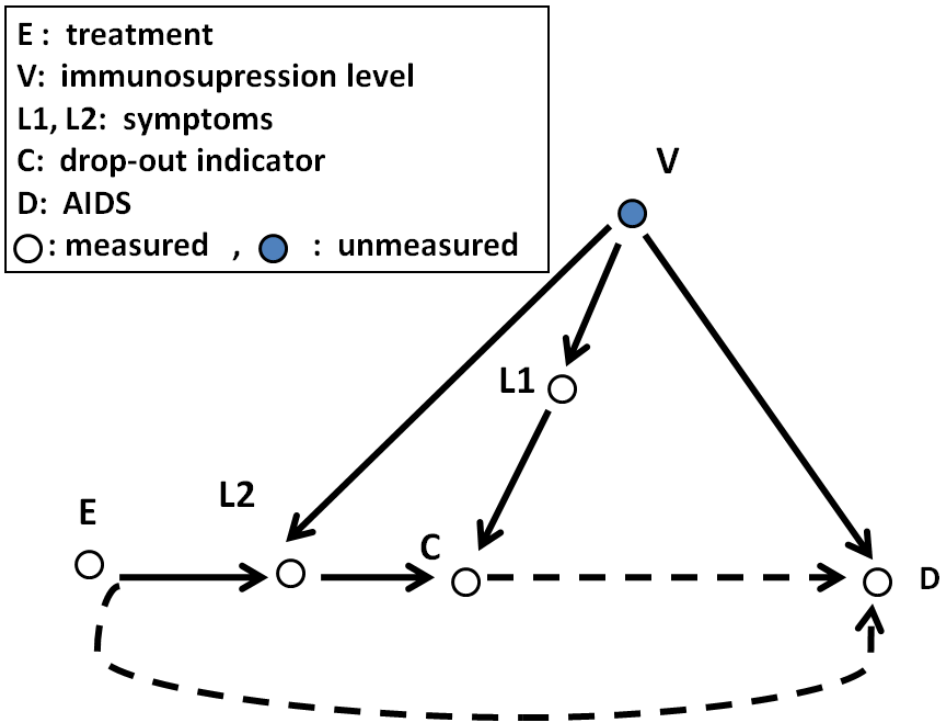
$$\frac{p_{c=0}\left(d,l,c'|e\right)}{p\left(d,l,c'|e\right)}=\frac{I_{\{0\}}\left(c'\right)}{\Pr\left(c=0|E=e,L=l\right)} \tag{12}$$

- we obtain

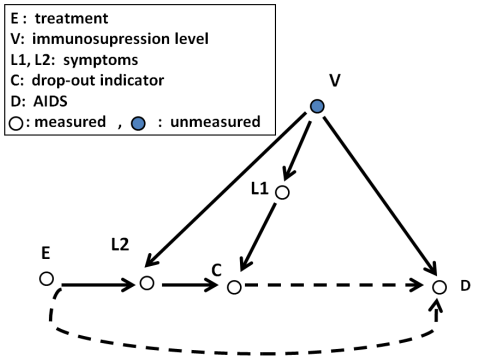
$$\underbrace{E_{c=0}\left(D|E=e\right)}_{\substack{\text{mean of } D \text{ given } E=e \\ \text{if nobody dropped out}}}=\underbrace{E\left\{\frac{I_{\{0\}}\left(C\right)}{\Pr\left(C=0|E=e,L\right)}D\middle|E=e\right\}}_{\text{the inverse probability weighted form}}$$

A more realistic example with drop-outs

- The preceding example is unrealistic because it assumed that the post-randomization side effects were not influenced by the patients' underlying immune status
- A more realistic DAG is

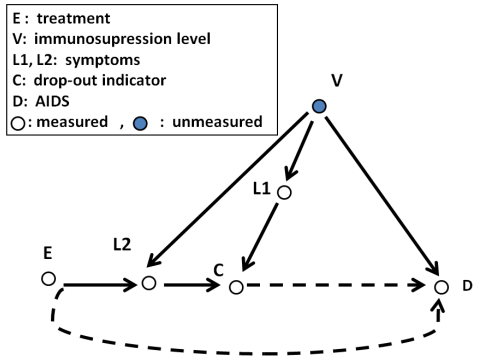


A more realistic example with drop-outs



- Even if (L_1, L_2) are measured we can't use the back-door formula for $p_{e,c=0}(d)$ because:
  1. (L_1, L_2) does not meet the back-door criterion because L_2 is a descendant of E
  2. L_1 does not meet the criterion because the path C, L_2, V, Y is unblocked by L_1
  3. L_2 does not meet the criterion because the path C, L_1, V, Y is unblocked by L_2

A more realistic example with drop-outs



- We will see later that $p_{e,c=0}(d)$ is identified and it holds that

$$p_{e,c=0}(d) = \sum_{l=(l_1,l_2)} p(d|e, c=0, l)\, p(l|e)$$

- But

$$p_{e,c=0}(d) \neq \sum_{l=(l_1,l_2)} p(d|e, c=0, l)\, p(l)$$

Section V: identifiability of the intervention law, the front-door adjustment and other results

- The front-door adjustment theorem
- Analysis of an example with two time dependent treatments
- Why regression analysis is wrong with time dependent treatments and covariates
- Identification theorem for time dependent treatment effects
- Back to our realistic drop-out example

Corollaries of the "do" calculus: the front-door adjustment

- **Definition:** In a DAG G a set of nodes Z satisfies *the front-door criterion relative to an ordered paired of nodes* (X, Y) iff:
  1. Z intercepts all directed paths between X and Y
  2. there is no back door path from X to Z, and
  3. all back door paths from Z to Y are blocked by X.
- **Theorem (Front door adjustment):** if in a DAG G, Z is a set of nodes that satisfies the front door criterion relative to the pair of nodes (X, Y) and if $p(x, z) > 0$ for all x, z, then

$$\boxed{p_x(y) = \sum_z p(z|x) \sum_{x'} p(y|x', z)\, p(x')}$$

Proof of the front-door adjustment theorem

$$
\begin{aligned}
p_x(y) &= \sum_z p_x(y|z)\, p_x(z) \\
&= \sum_z p_{x,z}(y)\, p_x(z) \;\; bc \;\; (Y \amalg Z|X)_{G_{\overline{X},\underline{Z}}} \;\; \textit{(by cdn 3)} \\
&= \sum_z p_z(y)\, p_x(z) \;\; bc \;\; (Y \amalg X|Z)_{G_{\overline{XZ}}} \;\; \textit{(by cdn 1)} \\
&= \sum_z p_z(y)\, p(z|x) \;\; bc \;\; (Z \amalg X)_{G_{\underline{X}}} \;\; \textit{(by cdn 2)} \\
&= \sum_z \left[\sum_{x'} p(y|x', z)\, p(x') \right] p(z|x) \;\; \textit{by cdn 3 and back-door adj.}
\end{aligned}
$$

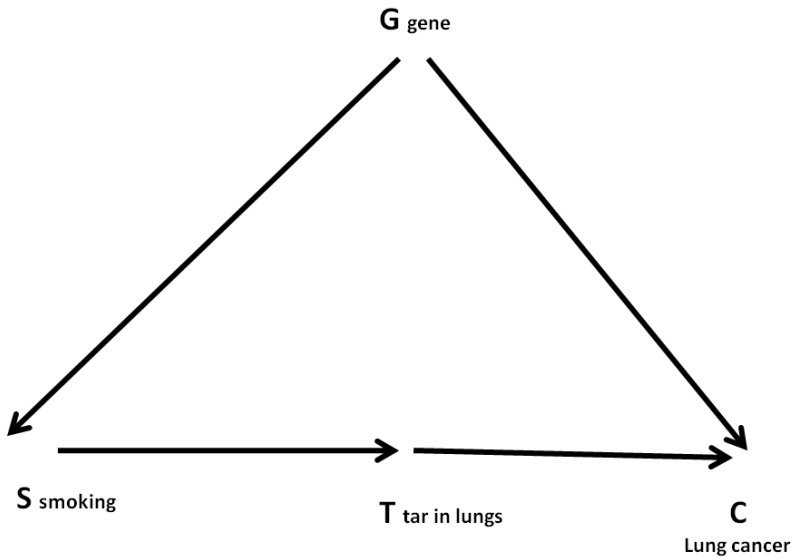
- Note: the second equality follows because condition 3 is $(Y \amalg Z|X)_{G_{\underline{Z}}}$ and this implies $(Y \amalg Z|X)_{G_{\overline{X},\underline{Z}}}$ because removing arcs in a DAG can not create new $d-$connections.

Intuition behind the front-door adjustment

- The intuition *(though not the proof)* of the front-door adjustment is as follows.
- Because by condition 1 the only directed paths between X and Y are paths that go through Z, then we can "decompose" the effect $p_x(y)$ in two parts:
  1. The effect of X on Z, i.e. $p_x(z)$
  2. The effect of Z on Y, i.e. $p_z(y)$
- Both $p_x(z)$ and $p_z(y)$ are identified:
  1. $p_x(z)$ is identified because by condition 2 there is no unblocked back door path between X and Z
  2. $p_z(y)$ is identified because by condition 3, X (which is measured) blocks all back door paths between Z and Y.

Example of the front-door adjustment theorem

Recall the example of smoking and lung cancer

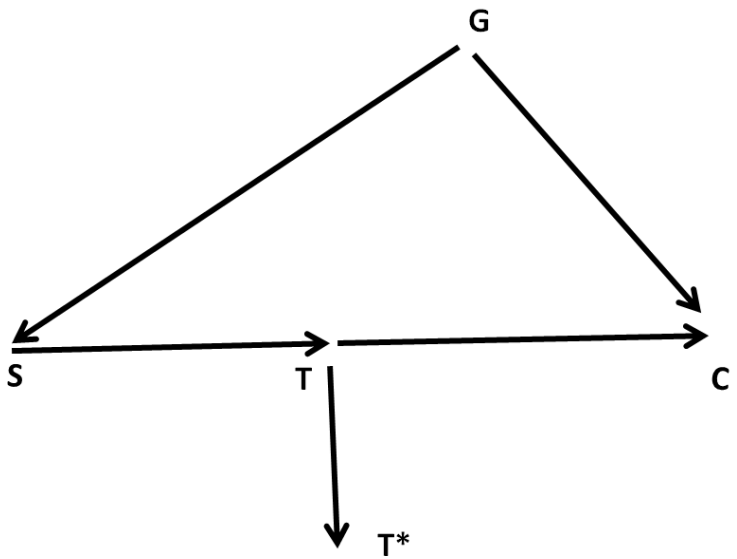


- T (tar) satisfies the front-door criterion relative to (S, C) hence

$$p_s(c) = \sum_z \left[\sum_{s'} p(c|s', t)\, p(s') \right] p(t|s)$$

Critiques to the example of smoking and lung cancer

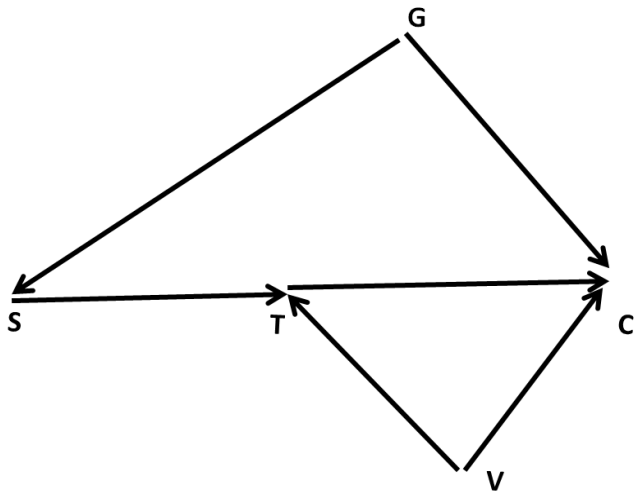
- **First critique:** *The causal model assumes that T is observed and measured with precision.*
- What if we actually measure T^* which is T plus some random error independent of everything?



- *T^* does not satisfy the front door condition* because condition 1 fails, T^* does not intercept all directed paths between S and C

Comments on the example of smoking and lung cancer

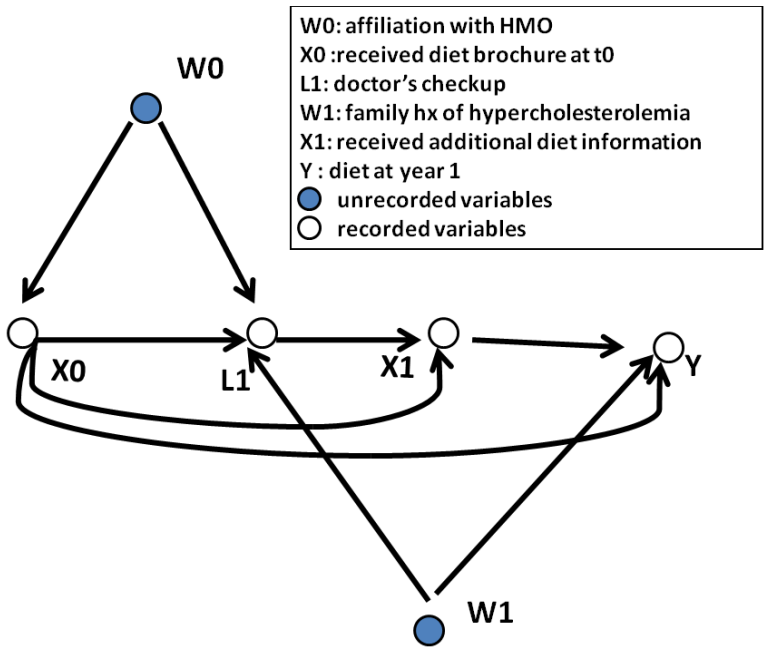
- **Second critique**: *the model assumes that the disturbances of T and C don't share common determinants*.
- But it is quite possible that there exist some biological factors V, e.g. a gene, that regulate both the way in which the lung stores tar and lung cancer



- *T does not satisfy the front door condition* because condition 3 fails, there are back-door paths between T and Y that are not blocked by V

Identification with time dependent treatments and covariates

- The following example illustrates the essential points of the situation that we consider next.



- We will see that even though both the front-door and the back-door criteria fail, $p_{x_1 x_2}(y)$ is identified

Observational study in DAG

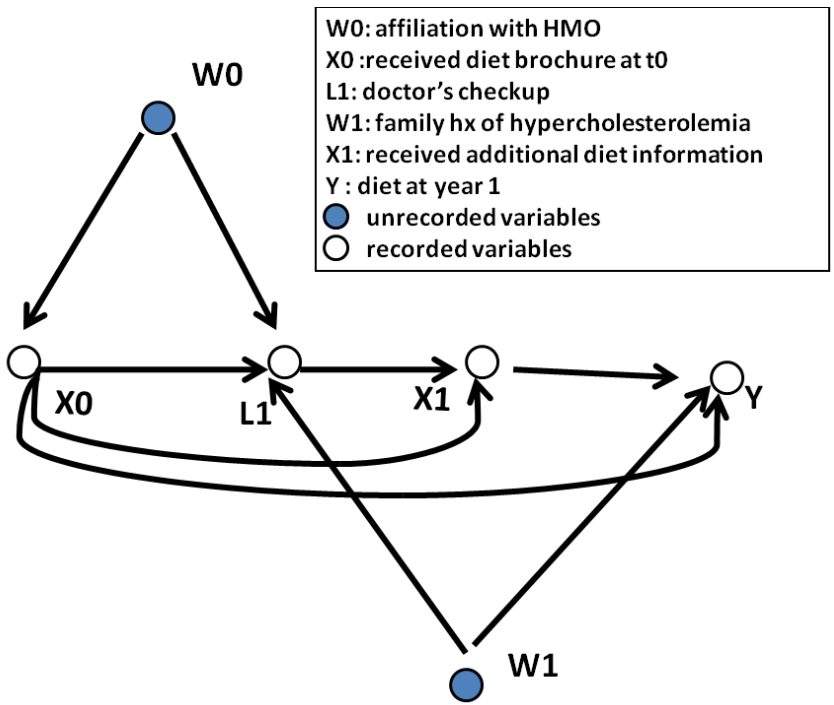
- As part of a national campaign on health diet awareness:
- At time t_0 the government
  1. distributes diet brochures at shopping malls
  2. encourages HMOs, through financial incentives, to mail diet brochures
- Six months later government distributes once again brochures at shopping malls
- One year later a survey asks
  1. Dietary habits (Y)
  2. Having received diet information at time t_0 (X_0)
  3. Having received any additional diet information later (X_1)
  4. Having had an annuals doctor's physical exam in the past year (L_1)
- **Objective**: to evaluate the impact of receiving different amounts of diet information on diet, i.e. $p_{x_0,x_1}(y)$
- **Unmeasured variables**
  1. Indicator of affiliation with an HMO (W_0)
  2. History of hypercholesterolemia in the family (W_1)

Arrows in the DAG of the example

1. Subjects in HMO's are more likely than gral population to
    1. receive diet brochure at time t_0 (arrow from W_0 to X_0)
    2. have an annual physical exam (arrow from W_0 to L_1)
2. Subjects with family history of hypercholesterolemia more like than gral population to
    1. have annual physical exam (arrow from W_1 to L_1)
    2. care about their diet (arrow from W_1 to Y)
3. HMO's brochures encourage annual check-ups (arrow from X_0 to L_1)
4. Patients that did not receive a brochure at t_0 are more likely than those that received it to care for a brochure six months later (arrow from X_0 to X_1)

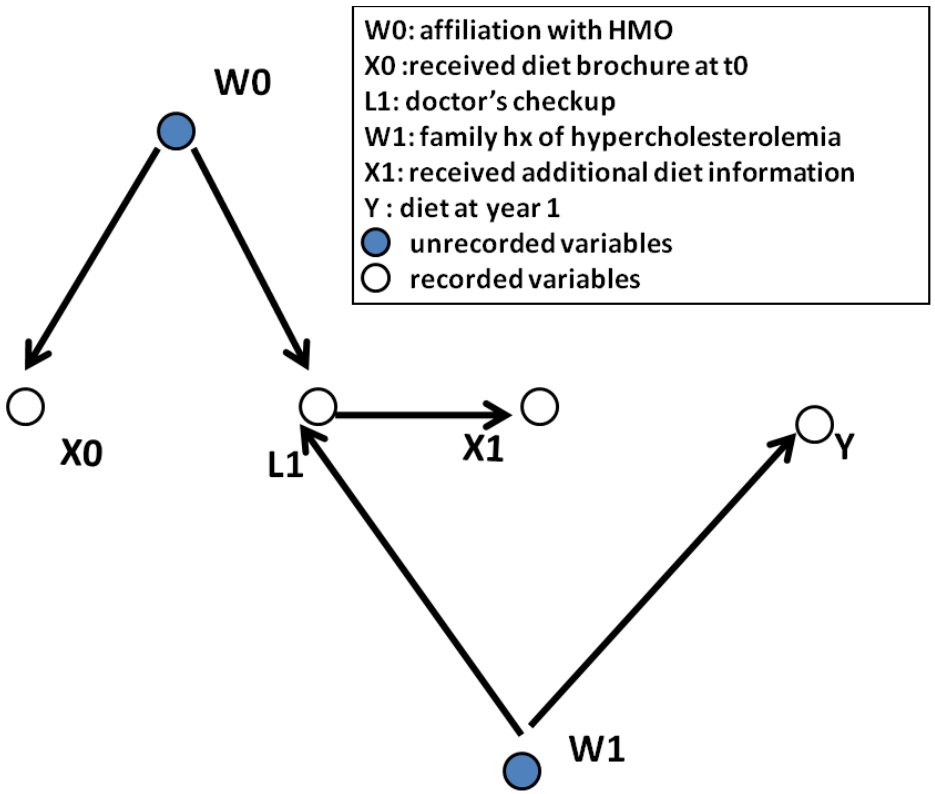
Front-door criterion not satisfied

- In our example, $X = (X_0, X_1)$. Will show that neither back-door nor front-door criteria are satisfied
- The **front door criterion fails** because there is no variable that intercepts all directed paths between X and Y.



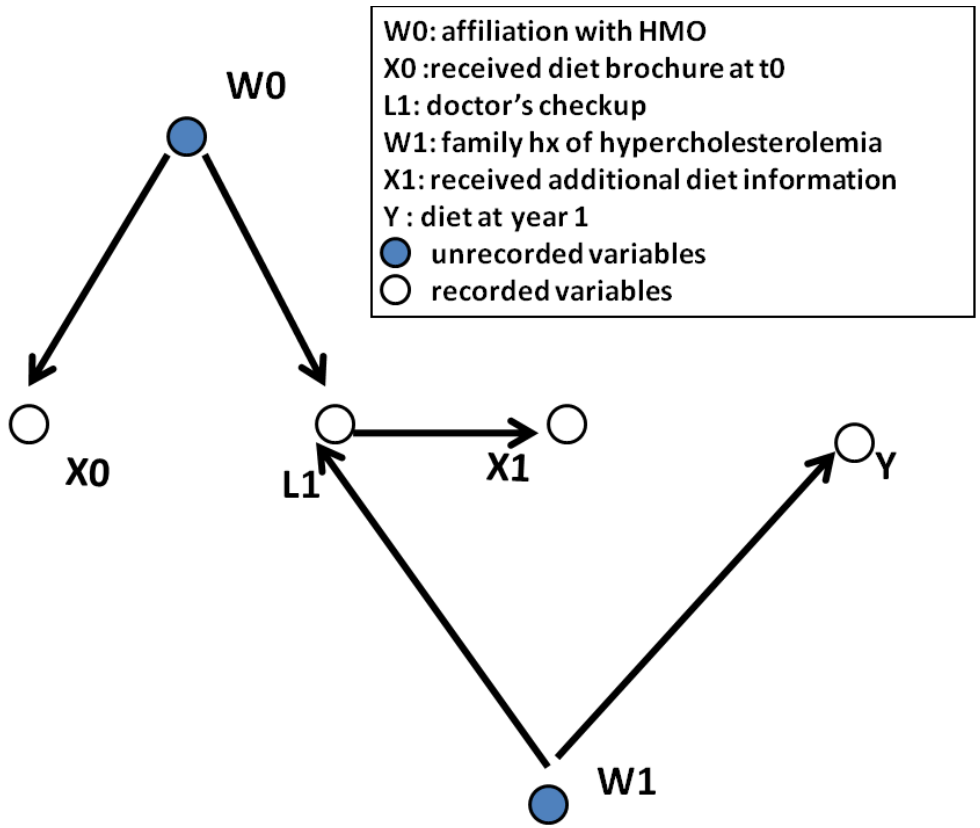
Back door criterion not satisfied

- Only two observed candidates for **back-door criterion** are $\varnothing$ and L_1
- $\varnothing$ does not satisfy the criterion because $(X \not\perp\!\!\!\perp Y)_{G_{\underline{X}_1,\underline{X}_0}}$
  - the path X_1, L_1, W_1, Y is unblocked in $G_{\underline{X}_1,\underline{X}_0}$



Back door criterion not satisfied

- $\{L_1\}$ does not satisfy the back-door criterion because $(X \not\perp\!\!\!\perp Y \mid L_1)_{G_{\underline{X}_1,\underline{X}_0}}$
  - the path X_0, W_0, L_1, W_1, Y is unblocked by L_1 in $G_{\underline{X}_1,\underline{X}_0}$



Identification of time dependent treatment effects

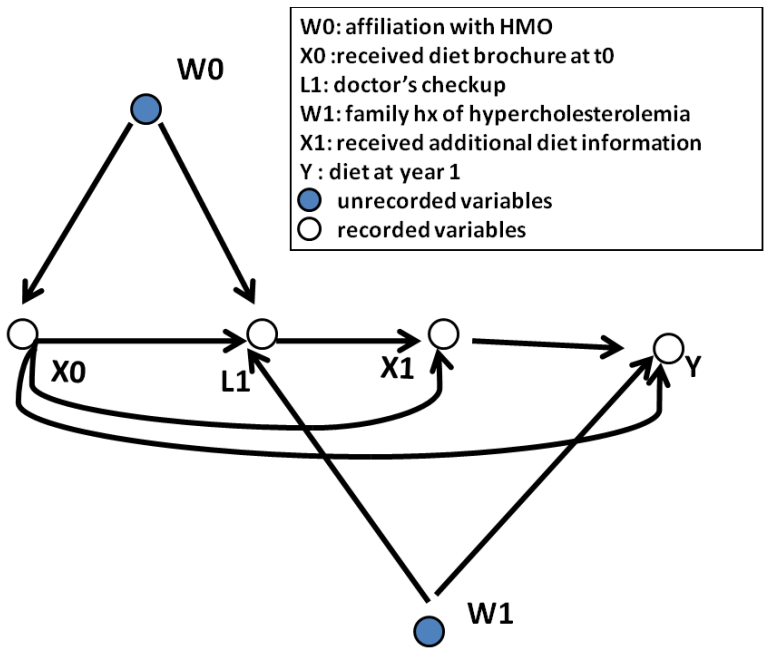
- **Result:** in the DAG of the example

$$p_{x_0,x_1}(y) = \sum_{l_1} p(y|l_1, x_0, x_1)\, p(l_1|x_0)$$

- **Corollary**:
  1. $p_{x_0,x_1}(y)$ depends only on the law of the measured variables $\{X_0, L_1, X_1, Y\}$.
  2. can estimate $p_{x_0,x_1}(y)$ consistently

Proof of result

$$
\begin{aligned}
p_{x_0,x_1}(y) &= p_{x_1}(y|x_0) && (\textit{rule 2}) \\
&= \sum_{l_1} p_{x_1}(y|l_1,x_0)\, p_{x_1}(l_1|x_0) \\
&= \sum_{l_1} p_{x_1}(y|l_1,x_0)\, p(l_1|x_0) && (\textit{rule 3}) \\
&= \sum_{l_1} p(y|l_1,x_0,x_1)\, p(l_1|x_0) && (\textit{rule 2})
\end{aligned}
$$



An interesting point

- We have seen that

$$p_{x_0,x_1}(y) = \sum_{l_1} p(y|l_1, x_0, x_1)\, p(l_1|x_0) \tag{13}$$

- However, it can be proved that $p_{x_0,x_1}(l_1)$ is not identified. This is essentially because with the measured variables we cannot block the back-door path X_0, W_0, L_1.

- (13) is the marginal distribution of Y under the *fictitious law* p^*

$$p^*(x_0', l_1, x_1', y) = p(y|l_1, x_0, x_1)\, I_{\{x_1\}}(x_1')\, p(l_1|x_0)\, I_{\{x_0\}}(x_0')$$

- This would be the intervention law if the causal DAG did not have the unmeasured covariates W_0 and W_1

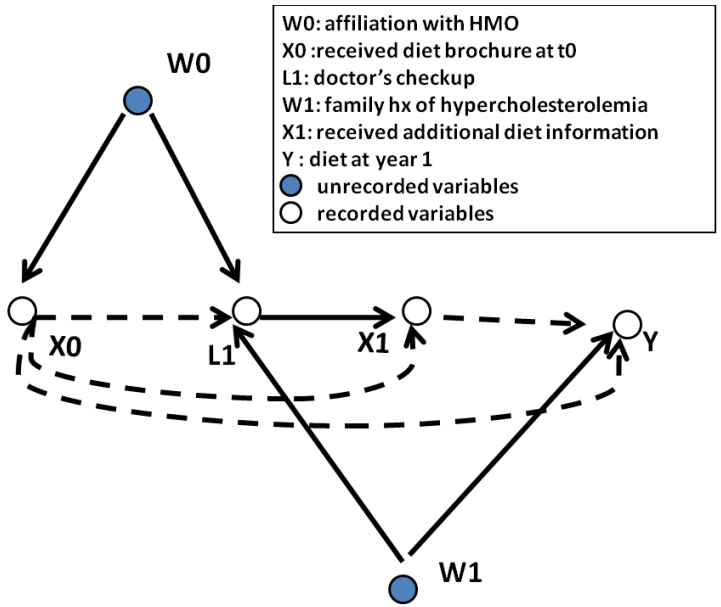
An interesting point

- We conclude that in this example

  1. we remove W_0 and W_1 from the DAG and compute the intervention law
  2. we use this *fictitious* intervention law to calculate the marginal distribution of Y. This gives the actual law of the counterfactual Y
  3. however, we cannot use this fictitious intervention law to compute the distribution of the counterfactual L

Why standard regression analysis is wrong

- I will now use our example to argue that regression analysis, whether adjusting or not for covariates, gives wrong answers.



- Suppose that neither X_0 nor X_1 have an effect on anything because, unknown to you, the dashed arrows are absent and consequently (by rule 3)

$$p_{x_0,x_1}(y) = p(y)$$

Why standard regression analysis is wrong

- Will a regression analysis tell you that (X_0, X_1) has no effect on Y?
- Besides X_0 and X_1 you also have in the database the covariate L_1
- So, your options are either to compute

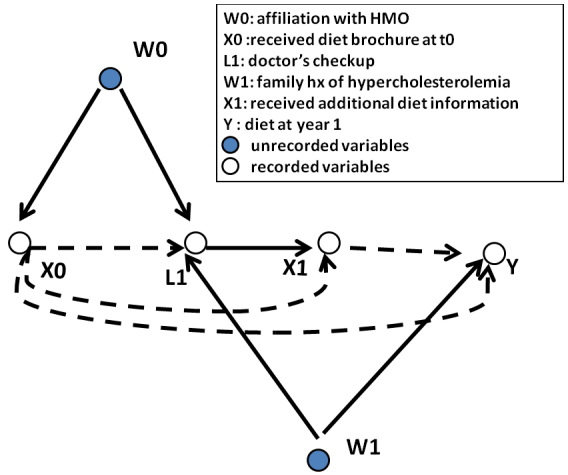
$$\boxed{p\left(y|x_0, x_1\right)\ (\textit{regression of } Y \textit{ on } X_0 \textit{ and } X_1)} \tag{14}$$

- or

$$\boxed{p\left(y|x_0, x_1, l_1\right)\ (\textit{regression of } Y \textit{ on } X_0, X_1 \textit{ and } L_1)} \tag{15}$$

Why standard regression analysis is wrong

- I will now show in the DAG that



even when the dashed arrows are absent, generally,

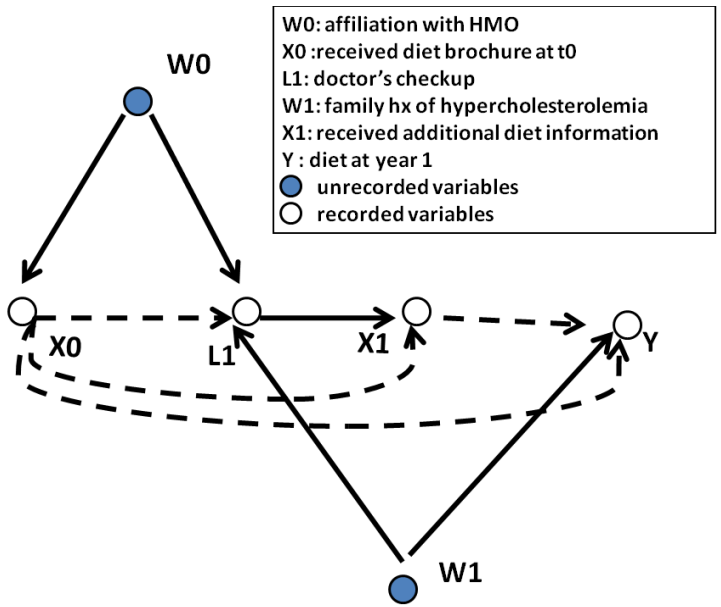
$$\boxed{p\left(y\middle|x_0, x_1\right) \text{ depends on } x_1}$$

and

$$\boxed{p\left(y\middle|x_0, x_1, l_1\right) \text{ depends on } x_0}$$

- So any option of regression analysis will lead you to falsely conclude that (X_0, X_1) has an effect on Y.

Why standard regression analysis is wrong

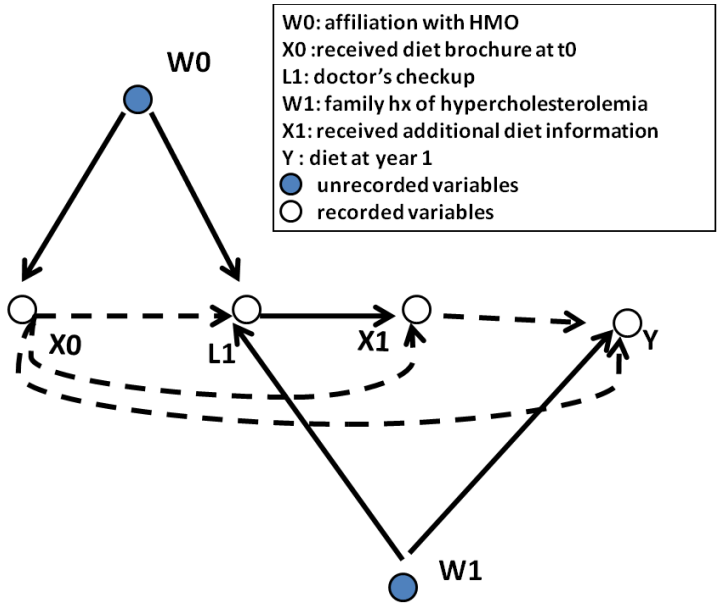


- $(X_1 \not\perp\!\!\!\perp Y)_G$ even if the dashed arrows are absent from G because the path Y, W_1, L_1, X_1 is unblocked.
- So, in general,

$$\boxed{p\left(y|x_0, x_1\right) \text{ depends on } x_1}$$

- **Key reason for failure**: by failing to condition on L_1, we do not block the back-door path X_1, L_1, W_1, Y

Why standard regression analysis is wrong



- $(X_0 \not\perp\!\!\!\perp Y \,|\, L_1)_G$ even if the dashed arrows are absent from G because the path Y, W_1, L_1, W_0, X_0 is unblocked by L_1
- So, in general,

$$\boxed{p\left(y \,|\, x_0, x_1, l_1\right) \text{ depends on } x_0}$$

- **Key reason for failure**: The pattern formed by the nodes X_0, W_0, L_1, W_1 and Y is an M structure. By conditioning on L_1 we generate *Berkson bias*

Why standard regression analysis is wrong

- **Conclusion**: in a longitudinal study, with a **time-dependent** covariate L_1 that
  1. is **associated** with previous exposure (X_0)
  2. is a **cause** of future exposure (X_1), and
  3. is **associated** with the outcome (Y)
- the coefficients of X_0 and X_1 in the either
  1. the regression of Y on (X_0, X_1) , or
  2. the regression of Y (X_0, X_1, L_1)
- **do not have a causal interpretation**.

Why standard regression analysis is wrong

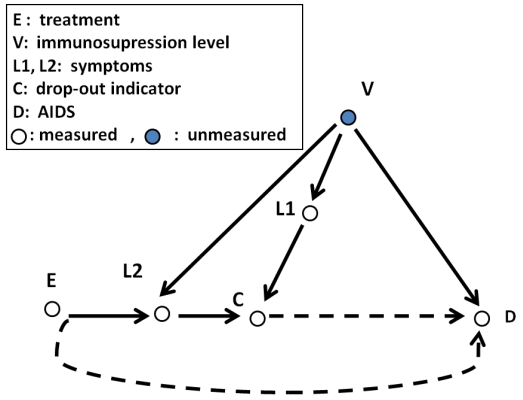
- This example shows that even in the *ideal world* absent of sampling variability or model misspecification, (so that conditional probabilities are known without sampling or model error)
  - a regression analysis which
    1. either does not adjust for the measured covariate L_1, or
    2. adjusts for the measured covariate L_1
  - can lead you to incorrectly conclude that (X_0, X_1) has an effect on Y
- The example also shows that even though regression analysis will give the wrong answers, the quantity of interest $p_{x_0,x_1}(y)$ is indeed a functional of the observed data law, i.e.

$$\boxed{p_{x_0,x_1}(y) = \sum_{l_1} p(y|l_1, x_0, x_1)\, p(l_1|x_0)}$$

- *You should check that if in the true DAG the dashed arrows are absent, then the expression on the RHS simplifies to* $p(y)$

Revisit our drop-out example

- We can now show the formula that identifies $p_{e,c=0}(d)$ in our DAG representing a realistic drop-out setting in a randomized trial



$$
\begin{aligned}
p_{e,c=0}(d) &= p_{c=0}(d|e) && (\textit{rule 2}) \\
&= \sum_{l=(l_1,l_2)} p_{c=0}(d|e,l)\, p_{c=0}(l_1|e) \\
&= \sum_{l=(l_1,l_2)} p_{c=0}(y|e,l)\, p(l|e) && (\textit{rule 3}) \\
&= \sum_{l=(l_1,l_2)} p(y|l,e,c=0)\, p(l|e) && (\textit{rule 2})
\end{aligned}
$$

Identification of time dependent treatment effects

- We will now give a Theorem (Pearl and Robins, 1995) that generalizes the preceding result.
- **Theorem:** let Y be a node in a causal DAG G that is disjoint with a set of nodes $X = \{X_0, ..., X_n\}$. Let N_k be the set of nodes that are non-descendants of $\{X_k, ..., X_n, Y\}$ in G. Suppose that $X_j \subset N_{j+1}$ for each $j \geq 0$, and that X_n is a non-descendant of Y. Let $X_{-1} = L_{-1} = \varnothing$. If there exists for each $j \geq 0$, a set of variables L_j such that
  1. $L_j \subset N_j$
  2. $\left(Y \amalg X_j | X_0, ..., X_{j-1}, L_0, ..., L_j\right)_{G_{\underline{X}_j, \overline{X}_{j+1}, ..., \overline{X}_n}}$
- then,

$$p_{x_0,...,x_n}(y) = \sum_{z_1,...,z_n} \left[p\left(y | l_0, ..., l_n, x_1, ..., x_n\right) \times \prod_{j=1}^{n} p\left(l_j | l_0, ..., l_{j-1}, x_1, ..., x_{j-1}\right) \right]$$

A super brief introduction to inference

- Non-parametric inference when the back-door criterion holds

- Methods for reducing dimension when the variables meeting the back-door criterion are high dimensional

1. Outcome regression adjustment
2. Propensity score regression adjustment
3. Stratification by the propensity score
4. Matching by the propensity score
5. Weighting by the inverse of the propensity score (known as inverse probability weighting, IPW)
6. Double-robust methods

- What is left?

Inference when the back door condition holds

- Rosembaun and Rubin **(JASA, 1984)** proved that when Z satisfies the back-door criterion for (X, Y), then the propensity score

$$\pi_x(Z) \equiv \Pr(X = x|Z)$$

also satisfies the back-door criterion for (X, Y)

- Then, if Z that satisfies the back-door criterion for (X, Y) .we have three forms of writing $E(Y_x)$,

$$\begin{aligned} E(Y_x) &= E\{E[Y|X = x, Z]\} \\ &= E\{E[Y|X = x, \pi_x(Z)]\} \\ &= E\left[\frac{I_{\{x\}}(X)}{\pi_x(Z)} Y\right] \end{aligned}$$

Non-parametric inference when the back-door condition holds

- The RHS of the equalities in the previous slide are three ways of writing the same functional of $p(x, y, z)$, and hence in particular, they agree at the empirical law
- Thus, we can estimate $E(Y_x)$ with

$$\widehat{E}(Y_x) = E_n\{E_n[Y|X = x, Z]\}$$

$$= E_n\{E_n[Y|X = x, \pi_{n,x}(Z)]\}$$

$$= E_n\left[\frac{I_{\{x\}}(X)}{\pi_{n,x}(Z)}Y\right]$$

where the subscript n indicates evaluation under the empirical law.

- **Big problem**: when Z is high dimensional, the estimator is unfeasible due to the **curse of dimensionality**

Methods for estimating causal expectations when Z is high dimensional

- To estimate the functional

$$\begin{aligned} E(Y_x) &= E\{E[Y|X=x,Z]\} \\ &= E\{E[Y|X=x,\pi_x(Z)]\} \\ &= E\left[\frac{I_{\{x\}}(X)}{\pi_x(Z)} I_{\{y\}}(Y)\right] \end{aligned}$$

  when Z is high dimensional we must reduce dimension by modeling one of the three choices

  1. $E[Y|X=x,Z]$
  2. $\pi_x(Z) \equiv \Pr(X=x|Z)$, or
  3. $\pi_x(Z) \equiv \Pr(X=x|Z)$ and $E[Y|X=x,\pi_x(Z)]$

- The different existing methods differ according to which of these choices they model.
- To be concrete, I will explain them for Y and X binary.

Methods for estimating causal expectations when Z is high dimensional

1. **Outcome regression adjustment**
2. Propensity score regression adjustment
3. Stratification by the propensity score
4. Matching by the propensity score
5. Weighting by the inverse of the propensity score (known as inverse probability weighting, IPW)
6. Double-robust methods

Outcome regression adjustment

- Outcome regression adjustment is based on the regression form

$$E\left(Y_x\right) = E\left\{E\left[Y|X = x, Z\right]\right\}$$

and it is essentially

$$\widehat{E}\left(Y_x\right) = E_n\left\{\widehat{E}\left[Y|X = x, Z\right]\right\}$$

ie.

$$\widehat{E}\left(Y_x\right) = n^{-1}\sum_{i=1}^{n}\widehat{E}\left[Y_i|X_i = x, Z_i\right]$$

where $\widehat{E}\left[Y|X = x, Z\right]$ is the fitted value from some parametric or semiparametric regression model for $E\left[Y|X = x, Z\right]$.

Algorithm for the outcome regression adjustment method

- Let $\lambda_i = P\left(Y_i = 1 | X_i, Z_i\right)$

1. We fit a logistic regression model of λ_i on A_i and L_i, for example

$$\log\left(\frac{\lambda_i}{1-\lambda_i}\right) = \beta_0 + \beta_1 X_i + \beta_2^T Z_i$$

   This is just an example! More complicated models with interactions and powers of the components of Z_i are allowed
2. We compute the fitted value

$$\widehat{\lambda}_i = \frac{e^{\widehat{\beta}_0 + \widehat{\beta}_1 x + \widehat{\beta}_2^T Z_i}}{1 + e^{\widehat{\beta}_0 + \widehat{\beta}_1 a + \widehat{\beta}_2^T Z_i}}$$

3. The outcome regression estimator of $P\left(Y_x = 1\right)$ (the causal risk for treatment x) is

$$\boxed{\widehat{e}_{x,R} = n^{-1} \sum_{i=1}^{n} \widehat{\lambda}_i}$$

Cautions about the outcome regression adjustment

- The logistic regression model is used to extrapolate the values of $\Pr\left(Y_i = 1 | X_i = x, Z_i\right)$ for subjects i that were not treated with $x \Rightarrow$
  1. If the logistic regression model is incorrect, then the method may yield biased estimators.
  2. But when Z is high dimensional it is quite possible that we may fail to specify a reasonably correct model!

- Because $\widehat{e}_{x,R}$ is a valid (i.e. consistent) estimator of $P\left(Y_x = 1\right)$, then a valid estimator of the causal odds ratio is

$$\frac{\widehat{e}_{1,R} / \left(1 - \widehat{e}_{1,R}\right)}{\widehat{e}_{0,R} / \left(1 - \widehat{e}_{0,R}\right)}$$

- A common mistake is to report as the regression adjusted estimator of the causal odds ratio, the value $\widehat{\beta}_1$.
- However,

$$\widehat{\beta}_1 \neq \frac{\widehat{e}_{1,R} / \left(1 - \widehat{e}_{1,R}\right)}{\widehat{e}_{0,R} / \left(1 - \widehat{e}_{0,R}\right)}$$

due to the lack of collapsibility of odds ratios.

Outcome regression adjustment with non-binary outcomes

- If the **outcomes** are continuous we may fit a **linear regression model, such as**

$$Y_i = \beta_0 + \beta_1 X_i + \beta_1^T Z_i + error_i$$

- Then, we estimate $E(Y_a)$, the causal average in treatment a with

$$\widehat{e}_{x,R} = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\beta}_0 + \widehat{\beta}_1 x + \widehat{\beta}_2^T Z_i\right)$$

- If, as in our example, the regression model does not include interactions with treatment, then the estimator of the so-called average treatment effect (ATE) $E(Y_1) - E(Y_0)$ is

$$\widehat{e}_{1,R} - \widehat{e}_{0,R}$$

- This is algebraically identical to $\widehat{\beta}_1$. This is why it is often said that the regression coefficient β_1 is the effect of X on Y **adjusted for confounding**

Methods for computing causal risks when L is high dimensional

1. Outcome regression adjustment
2. **Propensity score regression adjustment**
3. Stratification by the propensity score
4. Matching by the propensity score
5. Weighting by the inverse of the propensity score (known as inverse probability weighting, IPW)
6. Double-robust methods

Propensity score regression adjustment

- Propensity score regression adjustment is based on the form

$$E\left(Y_x\right) = E\left\{E\left[Y|X = x, \pi_x\left(Z\right)\right]\right\}$$

and it is essentially

$$\widehat{E}\left(Y_x\right) = E_n\left\{\widehat{E}\left[Y|X = x, \widehat{\pi}_x\left(Z\right)\right]\right\}$$

ie.

$$\widehat{E}\left(Y_x\right) = n^{-1}\sum_{i=1}^{n}\widehat{E}\left[Y_i|X_i = x, \widehat{\pi}_x\left(Z_i\right)\right]$$

where $\widehat{\pi}_x\left(Z_i\right)$ is a fitted value from a parametric or semiparametric logistic regression model for $\Pr\left(X = x|Z\right)$ and $\widehat{E}\left[Y|X = x, \widehat{\pi}_x\left(Z\right)\right]$ is the fitted value from some parametric or semiparametric model for $E\left[Y|X = x, \widehat{\pi}_x\left(Z\right)\right]$.

Propensity score regression adjustment

- The algorithm followed by the method of propensity score regression is:

1. We fit a logistic regression model for the propensity score, for example

$$\log\left\{\frac{\pi_1(Z_i)}{1-\pi_1(Z_i)}\right\} = \alpha_0 + \alpha_1^T Z_i$$

and compute the fitted values $\widehat{\pi}_i = e^{\widehat{\alpha}_0+\widehat{\alpha}_1^T Z_i} / \left(1 + e^{\widehat{\alpha}_0+\widehat{\alpha}_1^T Z_i}\right)$

2. With λ_i now denoting $\Pr(Y_i = 1 | X_i, \pi_1(Z_i))$, we fit another logistic regression model,

$$\log\left\{\frac{\lambda_i}{1-\lambda_i}\right\} = \beta_0 + \beta_1 X_i + \beta_2 \widehat{\pi}_i$$

and compute $\widehat{\lambda}_i = e^{\widehat{\beta}_0+\widehat{\beta}_1 x+\widehat{\beta}_2^T \widehat{\pi}_i} / \left(1 + e^{\widehat{\beta}_0+\widehat{\beta}_1 x+\widehat{\beta}_2^T \widehat{\pi}_i}\right)$

3. The estimator of $P(Y_x = 1)$, the risk for treatment x is

$$\boxed{\widehat{e}_{x,PS,REG} = n^{-1}\sum_{i=1}^{n}\widehat{\lambda}_i}$$

Caveat about the propensity score regression adjustment

- A problem with the propensity score regression adjustment method is that its validity relies on having **two models** correctly specified,
  1. one for the propensity score and
  2. another for the probability of the outcome
- If either model is wrong, then the method will yield biased estimators

Methods for computing causal risks when L is high dimensional

1. Outcome regression adjustment
2. Propensity score regression adjustment
3. **Stratification by the propensity score**
4. Matching by the propensity score
5. Weighting by the inverse of the propensity score (known as inverse probability weighting, IPW)
6. Double-robust methods

Stratification by the propensity score

- A simplification of the propensity score regression method, replaces the second regression with stratification by percentiles of the estimated propensity scores. The method works as follows
  1. Repeat step 1 of the preceding algorithm so as to compute the estimated prop. scores $\widehat{\pi}_i$
  2. Form, say five, strata according to the quintiles $\widehat{q}_j, j = 0, ..., 5$, of $\widehat{\pi}_i$ from the entire sample (treated and untreated) with $\widehat{q}_0 = 0$ and $\widehat{q}_5 = 1$
  3. Within each stratum, calculate the sample mean of Y_i for those treated with treatment x
  4. Estimate the risk $P(Y_x = 1)$ with the average of the five sample means obtained in step 3. That is,

$$\widehat{e}_{x,PS,SRAT} = \frac{1}{5}\sum_{j=1}^{5}\left\{\frac{1}{n_{x,j}}\sum_{\substack{i \text{ treated with } x \\ \text{and in strata } j}} Y_i\right\}$$

  where $n_{x,j}$ = number of subjects treated with x in the j^{th} stratum.

Iterative fitting of the propensity score model

- To fit the propensity score model Rosenbaum and Rubin (JASA, 1984) recommended that, following the formation of the strata (defined by, say, quintiles of the estimated prop. score) the analyst examine the degree of balance for each covariate in L within each stratum. Evidence of imbalance may reflect that the propensity score model is incorrect, and the need to iterate the model fitting with a refined propensity score model.

Caveats on the method of stratification by the propensity score

- Stratification by the propensity score is indeed a propensity score regression method with a special (quite restrictive) model for the outcome that assumes that
  - the mean of the outcome in each experimental group depends on the propensity score only through its quintile stratum.

- Most publications use stratification by quintiles owing to the recommendation of Rosembaum and Rubin, Biometrika, 1983, and JASA, 1984. It is often advocated that stratification by quintiles removes nearly 90% of the bias in the crude risks.
- However, in a simulation study reported in a recent article of Lunceford and Davidian (Statistics in Medicine, 2004) the method of stratification by quintiles of the prop. score showed substantially smaller gains in bias reduction.

Methods for computing causal risks when L is high dimensional

1. Outcome regression adjustment
2. Propensity score regression adjustment
3. Stratification by the propensity score
4. **Matching by the propensity score**
5. Weighting by the inverse of the propensity score (known as inverse probability weighting, IPW)
6. Double-robust methods

Propensity score matching

- Propensity score matching essentially relies of some form of non-parametric estimation of $E\left[Y|X=x,\widehat{\pi}_x\left(Z\right)\right]$ for some preliminary estimator of $\widehat{\pi}_x\left(Z\right)$
- The algorithm for propensity score matching is
  1. Compute $\widehat{\pi}_1\left(Z\right)$, the estimated propensity score for each subject, usually he fit from some parametric, e.g. logistic regression, model.
  2. Using some matching algorithm, e.g. nearest neighbor, kernel, etc
     1. Match each treated subject with, say k, untreated subjects (controls)
     2. Match each untreated subject with , say k, treated subjects.

Propensity score matching

- The matched propensity score estimates of $E\left(Y_{x=1}\right)$ and $E\left(Y_{x=0}\right)$ are

$$\widehat{e}_{1,PS,M} = \frac{1}{n}\left\{\sum_{\substack{i\text{:subject } i \\ \text{was treated}}} Y_i + \sum_{\substack{j\text{:subject } j \\ \text{was not treated}}} \overline{Y}_{T,j}\right\} \text{ and}$$

$$\widehat{e}_{0,PS,M} = \frac{1}{n}\left\{\sum_{\substack{j\text{:subject } j \\ \text{was not treated}}} Y_j + \sum_{\substack{i\text{:subject } i \\ \text{was treated}}} \overline{Y}_{c,i}\right\}$$

where

1. $\overline{Y}_{c,i}$ is the average of the outcomes for the matched controls for the i^{th} treated subject.
2. $\overline{Y}_{T,j}$ is the average of the outcomes for the matched treated subjects for the j^{th} control

Methods for computing causal risks when L is high dimensional

1. Outcome regression adjustment
2. Propensity score regression adjustment
3. Stratification by the propensity score
4. Matching by the propensity score
5. **Weighting by the inverse of the propensity score (known as inverse probability weighting, IPW)**
6. Double-robust methods

Inverse probability weighting

- IPW is based on the form

$$E\left(Y_x\right) = E\left[\frac{I_{\{x\}}\left(X\right)}{\pi_x\left(Z\right)}Y\right]$$

- It is computed as

$$\widehat{e}_{x,IPW} = \frac{\sum_{\substack{\text{all subjects } i \\ \text{with } X_i = x}} \frac{1}{\widehat{\pi}_{x,i}} Y_i}{\sum_{\substack{\text{all subjects } i \\ \text{with } X_i = x}} \frac{1}{\widehat{\pi}_{x,i}}}$$

Caveats about the IPW method

- The method relies on the propensity score model being right
  - it can give substantially biased results if the model is wrong because if so, each treated subject may misrepresent the right proportion of subjects in the population with the same prognostic factors.
- Even if the propensity score model is right, the estimator may have an undesirable behavior when the true propensity scores are close to 0 (for estimating risk if treated) and close to 1 (for estimating risk if untreated).
  - In most samples there will be nobody with $Z's$ corresponding to small propensity scores among the treated, so the estimator will be systematically over (or under)-estimating quite far from the truth if the estimated propensity scores are very close to 0 (or close to 1 if we are estimating the risk if untreated) because in such case some subjects may receive unduly large weights.
- It is because of the problem of unduly large weights that the method is not recommended when some estimated propensity scores are close to 0 or to 1.

Methods for computing causal risks when L is high dimensional

1. Outcome regression adjustment
2. Propensity score regression adjustment
3. Stratification by the propensity score
4. Matching by the propensity score
5. Weighting by the inverse of the propensity score (known as inverse probability weighting, IPW)
6. **Double-robust methods**

Double-robust methods

- We have seen two methods that rely on just one model being right:
  1. **Outcome regression adjustment:** relies on regression model for the **outcome** Y **given** A **and** L
  2. **IPW estimation:** relies on logistic regression model for the relationship between the **propensity score and** L
- Each method fails if the assumed models are misspecified.
- Double-robust (DR) methods are techniques that require that one specify both
  1. an outcome regression model
  2. a model for the propensity score
- But DR methods give valid inference if **one of the models is right, but not necessarily both!!!!**
- Contrast this with the method of **propensity score regression adjustment**. That method needed the specification of the same two models, but it **required that both models be correct** in order to give valid inferences

Double-robust methods

- Recall the outcome regression adjusted estimator

1. We fit a logistic regression model for $\lambda_i = \Pr(Y_i = 1|X_i, Z_i)$, for example
$$\log\left(\frac{\lambda_i}{1-\lambda_i}\right) = \beta_0 + \beta_1 X_i + \beta_2^T Z_i$$
2. We compute the fitted value
$$\widehat{\lambda}_i = \frac{e^{\widehat{\beta}_0 + \widehat{\beta}_1 x + \widehat{\beta}_2^T Z_i}}{1 + e^{\widehat{\beta}_0 + \widehat{\beta}_1 x + \widehat{\beta}_2^T Z_i}}$$
3. The outcome regression estimator of $P(Y_x = 1)$ (the risk for treatment x) is
$$\boxed{\widehat{e}_{x,R} = n^{-1}\sum_{i=1}^{n}\widehat{\lambda}_i}$$

Double-robust methods

- The double-robust estimator of $P\left(Y_a = 1\right)$ is computed by adding to the outcome regression estimator and **augmentation term**

$$\underbrace{\widehat{e}_{x,DR}}_{\text{DR estimator}} = \underbrace{\widehat{e}_{x,R}}_{\text{Outcome Reg Estimator}} + \underbrace{\widehat{d}_x}_{\text{Augmentation term}}$$

- Augmentation term definition

$$\widehat{d}_x = \frac{\sum_{\substack{\text{all subjects } i \\ \text{with } X_i = x}} \frac{1}{\widehat{\pi}_{x,i}} \left(Y_i - \widehat{\lambda}_i\right)}{\sum_{\substack{\text{all subjects } i \\ \text{with } X_i = x}} \frac{1}{\widehat{\pi}_{x,i}}}$$

- It can be shown that $\widehat{e}_{x,DR}$ is consistent for $E\left(Y_x\right)$ provided either the outcome regression model or the propensity score model is correct but not necessarily both

A brief tour for what we left...

- Inference for the causal effects of time dependent treatments in the presence of time dependent covariates
- Instrumental variables methods
- Principal stratum estimands
- Direct vs indirect effects
- Sensitivity analysis and best-worse case bounds for non-identified estimands
- Calculation of the probability of counterfactual statements.

Una invitacion...

- Si le ha interesado el curso, queda invitado al taller de causalidad que se realiza cada lunes de 19:15 a 21:30 hs en la Universidad Di Tella
- El taller es interdisciplinario y asisten al mismo economistas, epidemiologos y matematicos
- El taller es gratuito y abierto al publico en general
- Para mas informacion puede escribirme a arotnitzky@utdt.edu

APPENDIX: PROOF OF THE INDENTIFICATION THEOREM

Proof of the identification theorem

- **Proof:** We will show the absolute continuity by showing by induction that if $p_{x'}(v) > 0$ then $p(v_l|\overline{v}_{l-1}) > 0, l = 1, ..., k$. Suppose then that $p_{x'}(v) > 0$, then
  1. $p(v_1) > 0$ because
     1. if $v_1 \in x'$ then $p(v_1) > 0$ by (6) since PA_{V_1} is empty. and
     2. if $v_1 \notin x'$ then $p(v_1) = \Pr(f_1(U_1) = v_1) = p_{x'}(v_1)$ and consequently is true by the assumption $p_{x'}(v_1|) > 0$
  2. Suppose that $p(v_l|\overline{v}_{l-1}) > 0$ is true for $1, ..., j-1$, then it is true for $l = j$ because
     1. If $v_j \in x'$, then $p(v_j|\overline{v}_{j-1}) = p(x'_s|pa_s)$ for some s, and then $p(v_j|\overline{v}_{j-1}) > 0$ holds by (6)
     2. If $v_j \notin x'$, then by inductive assumption $p(\overline{v}_{j-1}) > 0$ and in such case, $p(v_j|\overline{v}_{j-1})$ is well defined and it holds that $p(v_j|\overline{v}_{j-1}) = \Pr(f(pa_j, U_j) = v_j) = p_{x'}(v_j|\overline{v}_{j-1}) > 0$

Proof of the identification theorem, continued

- Next,

$$
\begin{aligned}
& p_{x'}(v) \\
= & \left\{\Pi_{j=1}^{k} p_{x'}\left(v_j|pa_j\right)\right\} I_{\{p_{x'}(\cdot)>0\}}(v) && (16)\\
= & \left\{\Pi_{v_j \notin x'} p_{x'}\left(v_j|pa_j\right)\right\} I_{\{x'\}}(x)\, I_{\{p_{x'}(\cdot)>0\}}(v) && (17)\\
= & \left\{\Pi_{v_j \notin x'} \Pr\left(f_j\left(pa_j, U_j\right) = v_j\right)\right\} I_{\{x'\}}(x)\, I_{\{p_{x'}(\cdot)>0\}}(v) && (18)\\
= & \left\{\Pi_{v_j \notin x'} \Pr\left(f_j\left(pa_j, U_j\right) = v_j\right)\right\} I_{\{x'\}}(x)\, I_{\{p(\cdot)>0\}}(v) && (19)\\
= & \left\{\Pi_{v_j \notin x'} \Pr\left(f_j\left(pa_j, U_j\right) = v_j|PA_j = pa_j\right)\right\} I_{\{x'\}}(x)\, I_{\{p(\cdot)>0\}}(v) && (20)\\
= & \left\{\Pi_{v_j \notin x'} p\left(v_j|pa_j\right)\right\} I_{\{x'\}}(x)\, I_{\{p(\cdot)>0\}}(v) && (21)
\end{aligned}
$$

Proof of the identification theorem, continued

-
  1. (16) is true by the causal Markov condition
  2. (17) is true because $p_{x'}\left(x_s|pa_s\right) = I_{\{x'_s\}}\left(x_s\right)$
  3. (18) is true because $U_j \amalg \overline{V}_{j-1}\left(x'\right)$
  4. (19) is true because $I_{\{x'\}}\left(x\right) I_{\left\{p_{x'}(\cdot)>0\right\}}\left(v\right) = I_{\{x'\}}\left(x\right) I_{\{p(\cdot)>0\}}\left(v\right)$ since
     1. the left hand side equal 1 implies the right hand side equal 1 by absolute continuity of $p_{x'}\left(\cdot\right)$ with respect to $p\left(\cdot\right)$
     2. the right hand side equal 1 implies $x = x'$ and $p\left(v_j|pa_j\right) > 0$. But if $x = x'$, then $p\left(v_j|pa_j\right) = p_{x'}\left(v_j|pa_j\right)$ which shows that the left hand side is 1
  5. (20) is true because $U_j \amalg \overline{V}_{j-1}$ and because $\Pr\left(PA_j = pa_j\right) > 0$ and hence conditioning on $PA_j = pa_j$ is valid
  6. (21) is true by definition of $p\left(v_j|pa_j\right)$

References

- The following list of references is not comprehensive. There is a ton written about causal inference in longitudinal studies with time dependent treatments. I just give a brief list of papers at the end here, but you should go to Jamie Robins' web site for a comprehensive list.
- To read about causal diagrams I recommend that you read Judea Pearl's book (it is listed in the next slide.
- Also, go to his webpage at UCLA (type his name in google to find his page. He has tons of papers for downloading there.

Books

- Morgan, S. Winship, C.(2007). *Counterfactuals and Causal Inference.* Cambridge University Press. (a good introductory book)
- Manski, Ch. (1994). *Identification problems in social sciences* Harvard University Press. (causal modeling in econometrics and social sciences)
- Rubin, D. (2006) *Matched Sampling for Causal Effects*. Cambridge University Press (a collection of reprints of articles by the author)
- Pearl, J. (2000). *Causality: Models, Reasoning and Inference*. Cambridge University Press (a book about causal graphs)
- Rosenbaum, RP. (2002). *Observational Studies*, 2nd edn. New York: Springer-Verlag.

Books

- van der Laan MJ, Robins JM. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. Springer Verlag: New York (Advanced and very hard to read. It treats the theory for semiparametric models for causal inference)
- Tsiatis, A. (2006). *Semiparametric Theory and Missing Data.* Springer. (Treats the same theory as van der Laan and Robins, but at an introductory level. Only one chapter on causality, and only about point exposure studies).

The counterfactual model

- Rubin, DB. (1983). Estimating causal effects in randomized and non-randomized studies. *Journal of educational psychology*. 66, 688-701.2.
- Rubin, D., (1977), "Assignment to Treatment Group on the Basis of a Covariate," *Journal of Educational Statistics*, 2(1): 1-26.
- Rubin, D., (1978), "Bayesian inference for causal effects: The Role of Randomization", *Annals of Statistics,* 6: 34-58.
- Holland, P. (1986). Statistics and causal inference. *Journal of the American Statistical Association*. 81, 945-960.
- Hernan, M. (2004). A definition of causal effect for epidemiological research. *J Epidemiol Community Health*; 58:265–271.
- Crump, R., Hotz, V., Imbens, G. and Mitnik, O. (2006) Moving the Goalposts: Addressing Limited Overlap in Estimation of Average Treatment Effects by Changing the Estimand. Paper downloadable from ideas.repec.org/p/iza/izadps/dp2347.html (this paper has an extensive reference list)

Philosophycal issues around the definition of counterfactuals

- Robins JM, Greenland S. (2000). Comment on "causal inference without counterfactuals." *J Am Stat Assoc* 95:477–82.
- Greenland S. (2002) Causality theory for policy uses of epidemiologic measures. In: Murray CJ, Salomon JA, Mathers, CD, et al, eds. Summary measures of population health. Cambridge, MA: Harvard University Press/World Health Organization,
- Hernan, M. (2005). Invited Commentary: Hypothetical Interventions to Define Causal Effects— Afterthought or Prerequisite? *American Journal of Epidemiology.* 162. 618–620

Theory of propensity scores methods

- Rosenbaum, PR. and Rubin, DB. (1983). The Central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 41-55.
- Rosenbaum, PR. and Rubin, DB. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association.* 79, 516-524.
- Rosenbaum, PR and Rubin, D. (1985) The bias due to incomplete matching *Biometrics* 41:103-16
- Rosenbaum, PR and Rubin, D. (1985) Constructing a control group using multivariate matched sampling methods. American Statistician 39:33-8
- Rosenbaum, PR (1987) Model based direct adjustment. *Journal of the American Statistical Association*. 82, 387-94
- Rosenbaum, PR. (1998). Propensity score. In *Encyclopedia of Biostatistics*, Volume 5, Armitage P, Colton T (eds). Wiley: New York, 3551-3555.

Double-robust methodology

- Robins, J. and Rotnitzky, A. (2001). Comment on "Inference for semiparametric models: some questions and an answer', by Bickel and Kwon. *Statistica Sinica* 11:920-36. (this has the most up to date results on the theory of double robustness)
- Bang H, Robins J. (2005). Doubly robust estimation in Missing data and causal Inference Models. *Biometrics,* 61:692-972. (the best expository paper about double robustness at an expository level)
- Rotnitzky A, Faraggi D and Schisterman. Doubly robust estimation of the area under the receiver-operating characteristic curve in the presence of verification bias. Journal of the American Statistical Association, 2006; 101(475): 1276-1288. D (an application of double-robust methods to a problem not involving causality)

Double-robust methodology

- Tan, Z. (2006) A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association.* 101(476):1619-37. (connects double-robustness with non-parametric likelihood estimation)
- Kang, J. and Schafer, J. (2007) Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. (with discussion) *Statistical Science*. 523-539 (compares with other methods and criticizes double-robustness).

Surveys of causal inference methodology for point exposure studies

- Hernan, M. and Robins, J. (2006). Estimating causal effects from epidemiologic data. *J. Epidemiol. Community Health* 60;578-586. (discusses standardization and IPW methods)
- Lunceford, JK. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine* 23, 2937-2960. (compares prop. score stratification, regression and double-robust methods)
- D'Agostino RB. Propensity score methods for bias reduction in the comparison of treatment to a non-randomized control group. (1998) *Statistics in Medicine*; 17:2265 –2281. (discusses all methods but without derivations)
- Austin PC, Mamdani MM, Stukel TA, Anderson GM, Tu JV. (2005) The use of the propensity score for estimating treatment effects: administrative versus clinical data. *Statistics in Medicine* 24:1563–1578.

Surveys of causal inference methodology for point exposure studies

- Austin PC. (2008) A critical appraisal of propensity score matching in the medical literature 1996-2003 (provides an extensive list of papers in the medical literature where propensity score methodology was applied). *Statistics in Medicine,* 27. 2037-49.
- Austin PC, Mamdani MM. (2006). A comparison of propensity score methods: a case-study estimating the effectiveness of post-AMI statin use. *Statistics in Medicine* 2006; 25:2084–2106. (this paper has the Statin study discussed in these notes. Be aware that it inadequately implements stratification and matching by the propensity score because of problems of collapsibility explained in these notes)

Instrumental variables Just a few...

- Greenland, S. (2000) An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology.* 29, 722-729.
- Angrist, J. Imbens, G. and Rubin, D. (1996). Identification of causal effects using instrumental variables (with discussion). *J. of the American Statistical Association.* 91. 444-472.
- Angrist, J. and Pischke, J. S. (2008) *Mostly Harmless Econometrics: An Empiricist's Companion, Ch 4.*
- Hernan, M. and Robins, J. (2006) Instruments for Causal Inference, an epidemiologist dream? *Epidemiology* • Volume 17, Number 4, pp 360-372

Theory of causal inference with time dependent treatments **Why standard regression models don't work**. (http://www.biostat.harvard.edu/~robins/research.html).

- Robins JM. (1997). Causal Inference from Complex Longitudinal Data. *Latent Variable Modeling and Applications to Causality.* Lecture Notes in Statistics (120), M. Berkane, Editor. NY: Springer Verlag, pp. 69-117. (Good exposition of why standard regression models don't help with causal inference. Deals with G-computation algorithm and nested models but no marginal models.I recommend that you start with this article)
- Robins JM. (1986). A new approach to causal inference in mortality studies with sustained exposure periods - Application to control of the healthy worker survivor effect. Mathematical Modelling, 7:1393-1512.
- Robins JM. (1987). A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. Journal of Chronic Disease (40, Supplement), 2:139s-161s.

Theory of causal inference with time dependent treatments. **Marginal Structural Models.** (http://www.biostat.harvard.edu/~robins/research.html).

- Robins, J. (1998a). Marginal structural models. *In 1997 Proceedings of the American Statistical Association. American Statistical Association*, Alexandria, VA, 1–10.
- Robins, J. (1999a). Association, causation, and marginal structural models. Synthese 121, 151–179. MR1766776
- Robins, J. (1999b). Marginal structural models versus structural nested models as tools for causal inference. *Statistical Models in Epidemiology: The Environment and Clinical Trials.* Springer-Verlag, 95–134. MR1731682.

Theory of causal inference with time dependent treatments. **Marginal Structural Models.** (http://www.biostat.harvard.edu/~robins/research.html).

- Robins, J. (2000). Robust estimation in sequentially ignorable missing data and causal inference models. In Proceedings of the American Statistical Association Section on Bayesian Statistical Science 1999. American Statistical Association, Alexandria, VA, 6–10.
- Robins JM, Hernán M, Brumback B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550-560.

Theory of causal inference with time dependent treatments. **Structural Nested Models**. (http://www.biostat.harvard.edu/~robins/research.html).

- Robins, J. (1998b). Structural nested failure time models. *The Encyclopedia of Biostatistics.* John Wiley and Sons, Chichester, U.K., Chapter Survival Analysis, P.K. Andersen and N. Keidig (Section editors), 4372–4389.
- Robins JM, Blevins D, Ritter G, Wulfsohn M. (1992). G-estimation of the effect of prophylaxis therapy for pneumocystis carinii pneumonia on the survival of AIDS patients. Epidemiology, 3:319-33
- Robins JM. (1994). Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics*, 23:2379-2412.

Theory of causal inference with time dependent treatments. **Structural Nested Models**. (http://www.biostat.harvard.edu/~robins/research.html).

- Robins JM. (1997). Structural nested failure time models. In: Survival Analysis, P.K. Andersen and N. Keiding, Section Editors. *The Encyclopedia of Biostatistics*, P. Armitage and T. Colton, Editors. Chichester, UK: John Wiley & Sons, pp. 4372-4389.
- Robins JM, Rotnitzky A. (2004). Estimation of treatment effects in randomised trials with non-compliance and a dichotomous outcome using structural mean models. Biometrika 91: 763-783.

Data analysis using marginal structural models. (http://www.biostat.harvard.edu/~robins/research.html).

- Hernán M, Brumback B, Robins JM. (2000). Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology*, 11(5):561-570.
- Hernán M, Brumback B, Robins JM. (2001). Marginal structural models to estimate the joint causal effect of nonrandomized treatments. Journal of the American Statistical Association – Applications and Case Studies, 96(454):440-448.
- Hernán MA, Brumback B, Robins JM. (2002). Estimating the causal effect of zidovudine on CD4 count with a marginal structural model for repeated measures. *Statistics in Medicine,* 21:1689-1709.

Data analysis using structural nested models. (http://www.biostat.harvard.edu/~robins/research.html).

- Mark SD, Robins JM. (1993). Estimating the causal effect of smoking cessation in the presence of confounding factors using a rank preserving structural failure time model. *Statistics in Medicine*, 12:1605-1628.
- Witteman JC, d'Agostino RB, Stijnen T, Kannel WB, Cobb JC, deRidder MAJ, Hoffman A, Robins JM. (1998). G-estimation of causal effects: isolated systolic hypertension and cardiovascular death in the Framingham Study. *American Journal of Epidemiology*, 148:390-401.
- Hernán MA, Cole S, Margolick J, Cohen M, Robins J (2005). Structural accelerated failure time models for survival analysis in studies with time-varying treatments. *Pharmacoepidemiology and Drug Safety*. (Published online 19 Jan 2005)