A Modicum of Causal Inference Theory

Eric J Tchetgen Tchetgen

Departments of Biostatistics and Epidemiology,
Harvard University

GNS
5/1/2013

Introduction

There are generally two notions of causation:

1. Cause of an effect: first observe an event/outcome, and subsequently identify the causes or events that lead to the observed outcome.

Introduction

There are generally two notions of causation:

1. Cause of an effect: first observe an event/outcome, and subsequently identify the causes or events that lead to the observed outcome.
2. Effect of a cause: assess the effect of a well defined exposure or intervention. e.g. does smoking cause lung cancer? does AZT prevent the advent of AIDS among HIV infected patients?

Introduction

- An example of (1):In the 80s, when unusual high number of patients dying from a combination of syndromes including a rare Kaposi's skin cancer and pneumonia, the primary scientific objective at the time was to identify the cause of this outbreak. Eventually, HIV found to be the cause.

Introduction

- An example of (1):In the 80s, when unusual high number of patients dying from a combination of syndromes including a rare Kaposi's skin cancer and pneumonia, the primary scientific objective at the time was to identify the cause of this outbreak. Eventually, HIV found to be the cause.
- This series of lectures will focus on (2), as most common to biostats and epi methodological research and is easier to address as it does not require complete scientific understanding.

Introduction

- An example of (1):In the 80s, when unusual high number of patients dying from a combination of syndromes including a rare Kaposi's skin cancer and pneumonia, the primary scientific objective at the time was to identify the cause of this outbreak. Eventually, HIV found to be the cause.
- This series of lectures will focus on (2), as most common to biostats and epi methodological research and is easier to address as it does not require complete scientific understanding.
- It falls under the experimental paradigm, which is made explicit in the context of randomized experiments, but is still a useful paradigm when experiments cannot be performed for either practical or ethical reasons.

Introduction

- An example of (1):In the 80s, when unusual high number of patients dying from a combination of syndromes including a rare Kaposi's skin cancer and pneumonia, the primary scientific objective at the time was to identify the cause of this outbreak. Eventually, HIV found to be the cause.
- This series of lectures will focus on (2), as most common to biostats and epi methodological research and is easier to address as it does not require complete scientific understanding.
- It falls under the experimental paradigm, which is made explicit in the context of randomized experiments, but is still a useful paradigm when experiments cannot be performed for either practical or ethical reasons.
    - e.g. smoking and lung cancer.

Introduction

Why do we need a formal theory of causation?

- Makes explicit what we mean by "causal effect", that is what is the quantity/estimand we seek?

Introduction

Why do we need a formal theory of causation?

- Makes explicit what we mean by "causal effect", that is what is the quantity/estimand we seek?
- Explains the popular saying " association is not necessarily causation". therefore standard statistical methods may not be used to infere causation.

Introduction

Why do we need a formal theory of causation?

- Makes explicit what we mean by "causal effect", that is what is the quantity/estimand we seek?
- Explains the popular saying " association is not necessarily causation". therefore standard statistical methods may not be used to infere causation.
- Gives conditions under which "association is causation", therefore standard statistical methods may be used.

Introduction

Why do we need a formal theory of causation?

- Makes explicit what we mean by "causal effect", that is what is the quantity/estimand we seek?
- Explains the popular saying " association is not necessarily causation". therefore standard statistical methods may not be used to infere causation.
- Gives conditions under which "association is causation", therefore standard statistical methods may be used.
- Generally makes explicit assumptions needed for the identification of causal effects, and allows for the derivation of new statistical methods when standard and familiar methods fail.

Introduction

Our causal paradigm consists of :

- Defining causal quantities, this will be done in terms of counterfactuals.

Introduction

Our causal paradigm consists of :

- Defining causal quantities, this will be done in terms of counterfactuals.
- Stating assumptions necessary to identify causal quantities (nonparametrically).

Introduction

Our causal paradigm consists of :

- Defining causal quantities, this will be done in terms of counterfactuals.
- Stating assumptions necessary to identify causal quantities (nonparametrically).
- Defining a mathematical model to deal with the curse of dimensionality.

Introduction

Our causal paradigm consists of :

- Defining causal quantities, this will be done in terms of counterfactuals.
- Stating assumptions necessary to identify causal quantities (nonparametrically).
- Defining a mathematical model to deal with the curse of dimensionality.
- performing statistical inference which includes testing and estimating the magnitude of a causal effect given the observed data.

Part I:Causal Effects of a Point Exposure

Counterfactuals

- Suppose you are contemplating taking an aspirin for your headache, and the outcome Y denotes whether or not you are headache free within say the next hour.

Counterfactuals

- Suppose you are contemplating taking an aspirin for your headache, and the outcome Y denotes whether or not you are headache free within say the next hour.
- As a thought experiment, you may think of two potential outcome variables either of which may be observed depending on whether or not you decide to take the aspirin. That is:

$$\left\{ \begin{array}{c} Y_0 : \textit{headache outcome after not taking aspirin} \\ Y_1 : \textit{headache outcome after taking aspirin} \end{array} \right\}$$

Counterfactuals

- Y_a is the outcome that you would observe if possibly countering to fact you followed treatment $a \in \{0, 1\}$.

Counterfactuals

- Y_a is the outcome that you would observe if possibly countering to fact you followed treatment $a \in \{0, 1\}$.
- The english sentence: 'aspirin has no causal effect on my headache outcome $Y \Leftrightarrow$ a mathematical statement about my potential outcomes $Y_1 = Y_0$.

Counterfactuals

- Y_a is the outcome that you would observe if possibly countering to fact you followed treatment $a \in \{0, 1\}$.
- The english sentence: 'aspirin has no causal effect on my headache outcome $Y \Leftrightarrow$ a mathematical statement about my potential outcomes $Y_1 = Y_0$.
- Similarly, we can think of an individual with a beneficial causal effect of aspirin if $Y_1 > Y_0$, or one with a harmeful causal effect of aspirin $Y_1 < Y_0$.

Counterfactuals

- The fundamental problem of causal inference is that you only observe one of the two potential outcomes.

$$Y = AY_1 + (1 - A)Y_0$$

The outcome corresponding to the treatment you did indeed take. That is Y_A is the factual outcome, and Y_{1-A} is the counterfactual.

Counterfactuals

- The fundamental problem of causal inference is that you only observe one of the two potential outcomes.

$$Y = AY_1 + (1 - A)Y_0$$

  The outcome corresponding to the treatment you did indeed take. That is Y_A is the factual outcome, and Y_{1-A} is the counterfactual.
- So that if in the data sample, you happen to be a person with $A = 1$, we observe Y_1, and Y_0 is missing, and vice versa for a person with $A = 0$.

Counterfactuals

- The fundamental problem of causal inference is that you only observe one of the two potential outcomes.

$$Y = AY_1 + (1 - A)Y_0$$

  The outcome corresponding to the treatment you did indeed take. That is Y_A is the factual outcome, and Y_{1-A} is the counterfactual.
- So that if in the data sample, you happen to be a person with $A = 1$, we observe Y_1, and Y_0 is missing, and vice versa for a person with $A = 0$.
- Therefore, it is impossible to evaluate individual causal effects. This is fundamentally a missing data problem. The only difference is that the full data is never observed with probability one.

Counterfactuals

- However, all is not lost, as under some assumptions, we can still say something about population causal effects. For instance, consider the following finite population version of the previous headache example.

ID	Y_0	Y_1	$Y_1 - Y_0$
1	0	0	0
2	1	0	-1
3	0	1	1
4	1	0	-1
5	1	0	-1
6	0	1	1
7	1	0	-1
8	0	0	0

Counterfactuals

- A commonly used population causal effect is given by the average causal effect:

$$\psi = E(Y_1 - Y_0) = E(Y_1) - E(Y_0) = 1/4 - 1/2 = -1/4$$

Note that this estimand can be written as a functional of the two marginal distributions of Y_0, and Y_1, without requiring their joint distribution This is going to be key to identifying ψ.

Randomization

Identification through randomization: Suppose we randomize our population of patients with a headache to either aspirin or no aspirin with equal probability 1/2. So that the observed data corresponds to columns four and five of the following table and you don't see the first two columns.

ID	Y_0	Y_1	A	Y
1	0	0	0	0
2	1	0	1	0
3	0	1	0	0
4	1	0	1	0
5	1	0	0	1
6	0	1	1	1
7	1	0	0	1
8	0	0	1	0

Randomization

Now consider the following parameter based on the observed data

$$\tau = E\left(Y|A=1\right) - E\left(Y|A=0\right) = 1/4 - 2/4 = -1/4$$

so that in this population, the crude association τ between A and Y appears to coincide with the average causal effect of A on Y. We will formally prove this below.

Randomization

But first lets restate our assumptions.

- (CA) Consistency Assumption: $Y = Y_A$ w.p.1

Randomization

But first lets restate our assumptions.

- (CA) Consistency Assumption: $Y = Y_A$ w.p.1
- (RA) Randomization Assumption: $\{Y_0, Y_1\} \amalg A$

Randomization

But first lets restate our assumptions.

- (CA) Consistency Assumption: $Y = Y_A$ w.p.1
- (RA) Randomization Assumption: $\{Y_0, Y_1\} \amalg A$
- If CA and RA hold, then $\psi = E(Y_1) - E(Y_0) \overset{RA}{=}$
  $E(Y_1|A=1) - E(Y_0|A=0) \overset{CA}{=} E(Y|A=1) - E(Y|A=0) = \tau.$

Randomization

- Note that the randomization assumption is simply saying that since A is determine say by a coin flip, it should be completely independent of patients' pretreatment characteristics.

Randomization

- Note that the randomization assumption is simply saying that since A is determine say by a coin flip, it should be completely independent of patients' pretreatment characteristics.
- It helps to think of the potential outcomes $\{Y_0, Y_1\}$ as being underlying pretreatment latent variables that exist prior to the random treatment assignment, and therefore, should be unrelated to the latter.

Randomization

- Note however that this RA does not imply $Y \amalg A$ since by the consistency assumption $Y = AY_1 + (1 - A)Y_0$ is determined by treatment and therefore is a posttreatment variable.

Randomization

- Note however that this RA does not imply $Y \amalg A$ since by the consistency assumption $Y = AY_1 + (1 - A)Y_0$ is determined by treatment and therefore is a posttreatment variable.
- In fact, $Y \amalg A$ holds if and only if the null hypothesis $Y_1 \stackrel{D}{=} Y_0$ holds and is also implied by the so-called sharp null hypothesis $Y_1 = Y_0$ a.s.

Randomization

- The identification of ψ depends solely on the identification of the marginal means $E(Y_a)$.

Randomization

- The identification of ψ depends solely on the identification of the marginal means $E(Y_a)$.
- For continuous exposure A, it is easy to show that under (CA) and (RA*):

Randomization

- The identification of ψ depends solely on the identification of the marginal means $E(Y_a)$.
- For continuous exposure A, it is easy to show that under (CA) and (RA*):
  - $\{Y_a : a \in \text{supp}(A)\} \amalg A$,$E(Y_a) = E(Y|A = a)$.

Randomization

- The identification of ψ depends solely on the identification of the marginal means $E(Y_a)$.
- For continuous exposure A, it is easy to show that under (CA) and (RA*):
  - $\{Y_a : a \in \text{supp}(A)\} \amalg A$,$E(Y_a) = E(Y|A = a)$.
- So that the average causal effect of A and Y is identified by the crude association between A and Y.

Randomization

- The identification of ψ depends solely on the identification of the marginal means $E(Y_a)$.
- For continuous exposure A, it is easy to show that under (CA) and (RA*):
  - $\{Y_a : a \in \text{supp}(A)\} \amalg A$,$E(Y_a) = E(Y|A = a)$.
- So that the average causal effect of A and Y is identified by the crude association between A and Y.
- This gives a formal justification for using randomized studies to validly assess the effect of interventions. Note that we have assumed the absence of non-compliance, blinding and missing or censored data. We will return to these issues later.

Observational Study:The Case of Point exposure

- Suppose that randomization no longer holds, because the observed data comes from a point exposure/cross-sectional observational study, with observed data $\{L, A, Y\}$.

Observational Study:The Case of Point exposure

- Suppose that randomization no longer holds, because the observed data comes from a point exposure/cross-sectional observational study, with observed data $\{L, A, Y\}$.
- L is a rich vector of covariates that satisfies :
  (NUCA) No unmeasured confounding assumption holds:
  $\{Y_0, Y_1\} \amalg A|L$

Observational Study:The Case of Point exposure

- Suppose that randomization no longer holds, because the observed data comes from a point exposure/cross-sectional observational study, with observed data $\{L, A, Y\}$.
- L is a rich vector of covariates that satisfies :
  (NUCA) No unmeasured confounding assumption holds:
  $\{Y_0, Y_1\} \amalg A|L$
- Then we say that there are no unmeasured counfounders for the effect of A on Y.

Observational Study:The Case of Point exposure

- The intuition behind (NUCA) is similar to that of RA. Mainly, that we have measured enough covariates L, so that within levels of L, the data mimicks a randomized trial with the randomization probabilities now allowed to depend on L.

Observational Study:The Case of Point exposure

- The intuition behind (NUCA) is similar to that of RA. Mainly, that we have measured enough covariates L, so that within levels of L, the data mimicks a randomized trial with the randomization probabilities now allowed to depend on L.
- Conceptually, this can be achieved only if we are able to measure all common causes of A and Y (that is all risk factors for Y that also determine A).

Observational Study:The Case of Point exposure

- Next we show that the no unmeasured confounding assumption is sufficient to again identify $\{E(Y_a) : a\}$ and thus $\psi = E(Y_1) - E(Y_0)$ Without loss of generality, suppose L is categorical; then
  $E(Y_a) = E(E(Y_a|L)) = \sum_l E(Y_a|L=l) f_L(l)$
  $\overset{NUCA}{=} \sum_l E(Y_a|A=a, L=l) f_L(l)$
  $\overset{CA}{=} \sum_l E(Y|A=a, L=l) f_L(l)$
  $\equiv g(a)$

Observational Study:The Case of Point exposure

- $g(a)$ is known as the *direct standardization* of $E(Y|A=a,L)$. It is a special case of Robins' *G-formula* (which we will discuss in the longitudinal case).

Observational Study:The Case of Point exposure

- $g(a)$ is known as the *direct standardization* of $E(Y|A=a,L)$. It is a special case of Robins' *G-formula* (which we will discuss in the longitudinal case).
- Thus $\psi = g(1) - g(0) = \sum_l \{E(Y|A=1,L=l) - E(Y|A=0,L=l)\} f_L(l)$ is the *standardized risk difference*.

Observational Study:The Case of Point exposure

- $g(a)$ is known as the *direct standardization* of $E(Y|A=a,L)$. It is a special case of Robins' *G-formula* (which we will discuss in the longitudinal case).
- Thus $\psi = g(1) - g(0) =$
  $\sum_l \{E(Y|A=1,L=l) - E(Y|A=0,L=l)\} f_L(l)$ is the *standardized risk difference*.
- Under NUCA, we see that crude association$\neq$causation, as
  $\sum_l E(Y_a|A=a,L=l) f_L(l) = E(Y_a)$
  $\neq E(Y|A=a) = \sum_l E(Y|A=a,L=l) f_L(l|A=a)$.

Observational Study:The Case of Point exposure

- So that the crude risk difference does not have a causal interpretation. However, if NUCA holds, and either of the following conditions holds:

$$Y \amalg L|A \text{ or } A \amalg L \tag{1}$$

then $E(Y_a) = E(Y|A = a)$ and L is a non-confounder, so that this implies that RA actually holds.

Observational Study:The Case of Point exposure

Proof:

- In the first case,
  $E(Y_a) = \sum_l E(Y|A=a, L=l) f_L(l)$
  $= \sum_l E(Y|A=a) f_L(l) = E(Y|A=a)$;

Observational Study:The Case of Point exposure

Proof:

- In the first case,
  $E(Y_a) = \sum_l E(Y|A=a, L=l) f_L(l)$
  $= \sum_l E(Y|A=a) f_L(l) = E(Y|A=a)$;
- In the second case,
  $E(Y_a) = \sum_l E(Y|A=a, L=l) f_L(l)$
  $= \sum_l E(Y|A=a, L=l) f_L(l|A=a) = E(Y|A=a)$

Observational Study:The Case of Point exposure

- In general,the point exposure G-formula is written

$$E\left(Y_a\right) = \int E\left(Y|A = a, L = l\right) dF(l)$$

Observational Study:The Case of Point exposure

- In general,the point exposure G-formula is written

$$E\left(Y_a\right) = \int E\left(Y|A = a, L = l\right) dF(l)$$

- The left-hand side is the mean of a counterfactual (latent variable), the right–hand side is a functional of the observed data, which is always well defined but only has a causal interpretation under the no unmeasured assumption.

Observational Study:The Case of Point exposure

- This functional is not a conditional expectation of the observed data, therefore, cannot be estimated directly such as the crude. However, the pluggin principle may be used as discussed below.

Observational Study:The Case of Point exposure

- This functional is not a conditional expectation of the observed data, therefore, cannot be estimated directly such as the crude. However, the pluggin principle may be used as discussed below.
- The G-formula is not restricted to the mean,other versions:

Observational Study:The Case of Point exposure

- This functional is not a conditional expectation of the observed data, therefore, cannot be estimated directly such as the crude. However, the pluggin principle may be used as discussed below.
- The G-formula is not restricted to the mean,other versions:
  - $E\left(Y_a|V=v\right)=\int E\left(Y|A=a,L=l\right)dF(l|V=v)$, where V is contained in L

Observational Study:The Case of Point exposure

- This functional is not a conditional expectation of the observed data, therefore, cannot be estimated directly such as the crude. However, the pluggin principle may be used as discussed below.
- The G-formula is not restricted to the mean,other versions:
  - $E\left(Y_a|V=v\right)=\int E\left(Y|A=a,L=l\right)dF(l|V=v)$, where V is contained in L
  - $f\left(Y_a|V=v\right)=\int f\left(Y|A=a,L=l\right)dF(l|V=v)$.

G-computation

- Given the observed data $O_i = (Y_i, A_i, L_i)$, G-computation generally refers to nonparametric inference on the G-formula $g(a) = \sum_l E(Y|A = a, L = l) f_L(l)$.

G-computation

- Given the observed data $O_i = (Y_i, A_i, L_i)$, G-computation generally refers to nonparametric inference on the G-formula $g(a) = \sum_l E(Y|A = a, L = l) f_L(l)$.
- A natural nonparametric estimator of $g(a)$ is given by the nonparametric pluggin estimator, which requires nonparametric estimates of $E(Y|A = a, L = l) = b(a, l)$ written $\widehat{b}(a, l)$ and of $F_L(l)$ written $\widehat{F}_L(l)$.

G-computation

- Until otherwise stated, assume both A and L are categorical variables with low to moderate number of levels, so that $\widehat{b}(a, l)$ is given by the stratified sample average:

$\widehat{b}(a, l) = \sum_{i=1}^{n} I\left(A_i = a, L_i = l\right) Y_i / \sum_{i=1}^{n} I\left(A_i = a, L_i = l\right)$
and $\widehat{f}_L\left(l\right) = n^{-1} \sum_{i=1}^{n} I\left(L_i = l\right)$

G-computation

- Until otherwise stated, assume both A and L are categorical variables with low to moderate number of levels, so that $\widehat{b}(a,l)$ is given by the stratified sample average:

$\widehat{b}(a,l) = \sum_{i=1}^{n} I\left(A_i = a, L_i = l\right) Y_i / \sum_{i=1}^{n} I\left(A_i = a, L_i = l\right)$
and $\widehat{f}_L\left(l\right) = n^{-1} \sum_{i=1}^{n} I\left(L_i = l\right)$

- The nonparametric estimator of the G-formula is given by:

$$\begin{aligned}
\widehat{g}(a) &= \sum_{l} \widehat{b}(a,l) \widehat{f}_L\left(l\right) = \sum_{l} \widehat{b}(a,l) n^{-1} \sum_{i=1}^{n} I\left(L_i = l\right) \\
&= n^{-1} \sum_{i=1}^{n} \sum_{l} \widehat{b}(a,l) I\left(L_i = l\right) = n^{-1} \sum_{i=1}^{n} \widehat{b}(a, L_i)
\end{aligned}$$

G-computation Asymptotic Distribution

One can show that

$$
\begin{aligned}
& n^{1/2}\left(\widehat{g}(a)-g\left(a\right)\right) \\
= \quad & n^{-1/2}\sum_{i=1}^{n}\left\{\frac{I\left(A_i=a\right)}{f_{A|L}\left(A_i|L_i\right)}\left(Y_i-b(A_i,L_i)\right)+b(a,L_i)-g\left(a\right)\right\}+o_p(1) \\
= \quad & n^{-1/2}\sum_{i=1}^{n}IF_i\left(a\right)+o_p(1) \\
\sim \quad & N\left(0,E\left(IF_i\left(a\right)^2\right)\right)
\end{aligned}
$$

G-computation:Mathematical Interlude

A wald type 95%CI for $\psi = g(1) - g(0)$ is given by:

$$\widehat{g}(1) - \widehat{g}(0) \pm 1.96 \sqrt{n^{-1} \sum_i \left(\widehat{IF}_i(0) - \widehat{IF}_i(1) \right)^2}$$

where $\widehat{IF}_i(a) = \left\{ \frac{I(A_i=a)}{\widehat{f}_{A|L}(a|L_i)} \left(Y_i - \widehat{b}(A_i, L_i) \right) + \widehat{b}(a, L_i) - \widehat{g}(a) \right\}$,
and $\widehat{f}_{A|L}(a|L_i = l) = \frac{\sum_i I(A_i=a, L_i=l)}{\sum_i I(L_i=l)}$ is the nonparametric estimator of $f_{A|L}(a|L_i = l)$,the probability of receiving treatment $A = a$ given $L = l$.

G-computation:Mathematical Interlude

- The term $b(a, L_i) - g(a)$ reflects the variability due to the estimation of $F_L(l)$, whereas $\frac{I(A_i=a)}{f_{A|L}(a|L_i)}(Y_i - b(a, L_i))$ captures the variability due to the estimation of $b(a, L_i)$.

G-computation:Mathematical Interlude

- The term $b(a, L_i) - g(a)$ reflects the variability due to the estimation of $F_L(l)$, whereas $\frac{I(A_i=a)}{f_{A|L}(a|L_i)}(Y_i - b(a, L_i))$ captures the variability due to the estimation of $b(a, L_i)$.
- It is interesting to note that the influence function and thus the variance of $\widehat{\psi} = \widehat{g}(1) - \widehat{g}(0)$ depends on the treatment process $\left\{f_{A|L}(a|L_i) : a\right\}$,even though the pluggin estimator described above appears not to.

G-computation:Mathematical Interlude

- The term $b(a, L_i) - g(a)$ reflects the variability due to the estimation of $F_L(l)$, whereas $\frac{I(A_i=a)}{f_{A|L}(a|L_i)}(Y_i - b(a, L_i))$ captures the variability due to the estimation of $b(a, L_i)$.
- It is interesting to note that the influence function and thus the variance of $\widehat{\psi} = \widehat{g}(1) - \widehat{g}(0)$ depends on the treatment process $\left\{f_{A|L}(a|L_i) : a\right\}$,even though the pluggin estimator described above appears not to.
- In fact, we give a different representation of the nonparametric estimator of the G-formula which makes explicit its dependence on the estimated treatment process.

G-computation:dual representation

- It can be shown that the nonparametric G-computation estimator $\widehat{g}(1) - \widehat{g}(0)$ has a dual representation as an inverse-probability of treatment weighted (iptw) estimator:

G-computation:dual representation

- It can be shown that the nonparametric G-computation estimator $\widehat{g}(1) - \widehat{g}(0)$ has a dual representation as an inverse-probability of treatment weighted (iptw) estimator:
- $\widehat{g}(a) = \sum_l \widehat{b}(a, l)\widehat{f}_L(l) = \frac{\sum_{s=1}^{n} I(A_s{=}a) Y_s \widehat{f}^{-1}_{A|L}(A_s|L_s)}{\sum_{s=1}^{n} I(A_s{=}a) \widehat{f}^{-1}_{A|L}(A_s|L_s)}$

G-computation:dual representation

- $\widehat{f}_{A|L}\left(a|L=l\right)=\frac{\sum_i I(A_i=a,L_i=l)}{\sum_i I(L_i=l)}$ is the nonparametric estimator of the treatment conditional probability mass function $f_{A|L}\left(a|L=l\right)$.

G-computation:dual representation

- $\widehat{f}_{A|L}\left(a|L=l\right)=\frac{\sum_i I(A_i=a,L_i=l)}{\sum_i I(L_i=l)}$ is the nonparametric estimator of the treatment conditional probability mass function $f_{A|L}\left(a|L=l\right)$.
- So that the nonparametric G-formula estimator has a (exact) dual representation as an inverse-probability-of-treatment weighted (iptw) estimator, and this also shows that the latter is asymptotically efficient.

G-computation:dual representation

- One can easily show (try it) that the iptw estimator $\widehat{g}(a)$ is the solution to the estimating equation:$\sum_{i=1}^{n} \frac{I(A_i=a)}{\widehat{f}_{A|L}(a|L_i=l)} (Y_i - \widehat{g}(a)) = 0.$

G-computation:dual representation

- One can easily show (try it) that the iptw estimator $\widehat{g}(a)$ is the solution to the estimating equation:$\sum_{i=1}^{n} \frac{I(A_i=a)}{\widehat{f}_{A|L}(a|L_i=l)} (Y_i - \widehat{g}(a)) = 0$.
- In contrast to the estimating equation for the conditional mean $\mu(a) = E(Y|A=a)$ given by: $\sum_{i=1}^{n} I(A_i = a)(Y_i - \widehat{\mu}(a)) = 0$.

G-computation:dual representation

- One can easily show (try it) that the iptw estimator $\widehat{g}(a)$ is the solution to the estimating equation:$\sum_{i=1}^{n} \frac{I(A_i=a)}{\widehat{f}_{A|L}(a|L_i=l)} (Y_i - \widehat{g}(a)) = 0$.
- In contrast to the estimating equation for the conditional mean $\mu(a) = E(Y|A=a)$ given by: $\sum_{i=1}^{n} I(A_i = a)(Y_i - \widehat{\mu}(a)) = 0$.
- We again see that if $f_{A|L}(a|L_i = l) = f_A(a)$, then $E(Y|A=a) = E(Y_a)$ and association is causation.

G-computation:dual representation

- One can easily show (try it) that the iptw estimator $\widehat{g}(a)$ is the solution to the estimating equation:$\sum_{i=1}^{n} \frac{I(A_i=a)}{\widehat{f}_{A|L}(a|L_i=l)} (Y_i - \widehat{g}(a)) = 0$.
- In contrast to the estimating equation for the conditional mean $\mu(a) = E(Y|A=a)$ given by: $\sum_{i=1}^{n} I(A_i = a)(Y_i - \widehat{\mu}(a)) = 0$.
- We again see that if $f_{A|L}(a|L_i = l) = f_A(a)$, then $E(Y|A=a) = E(Y_a)$ and association is causation.
- If $f_{A|L}(a|L_i = l) \neq f_A(a)$ and L is a risk factor of Y, then to obtain a counterfactual mean $E(Y_a)$, the recipe is to weight the estimating function for the conditional mean by $\widehat{f}^{-1}_{A|L}(a|L_i = l)$ to adjust for confounding by L.

G-computation:dual representation

- Why does weigthing adjust for confounding?

G-computation:dual representation

- Why does weigthing adjust for confounding?
- Essentially, creates a 'pseudo-population' in which A is no longer associated with Y; so that the crude mean in this population has a counterfactual interpretation.

G-computation:dual representation

- Why does weigthing adjust for confounding?
- Essentially, creates a 'pseudo-population' in which A is no longer associated with Y; so that the crude mean in this population has a counterfactual interpretation.
- For example, suppose L is binary and we observe :

G-computation:dual representation

- Why does weigthing adjust for confounding?
- Essentially, creates a 'pseudo-population' in which A is no longer associated with Y; so that the crude mean in this population has a counterfactual interpretation.
- For example, suppose L is binary and we observe :

N	A	L	$E(Y\|A=a, L=l)$
4000	1	0	24
3000	1	1	36
8000	0	0	10
9000	0	1	22

G-computation:dual representation

- Why does weigthing adjust for confounding?
- Essentially, creates a 'pseudo-population' in which A is no longer associated with Y; so that the crude mean in this population has a counterfactual interpretation.
- For example, suppose L is binary and we observe :

N	A	L	$E(Y\|A=a, L=l)$
4000	1	0	24
3000	1	1	36
8000	0	0	10
9000	0	1	22

- So that $f_{A|L}(A=1|L=1) = 1/4, f_{A|L}(A=1|L=0) = 1/3$ and thus L predicts A.

G-computation:dual representation

- Why does weigthing adjust for confounding?
- Essentially, creates a 'pseudo-population' in which A is no longer associated with Y; so that the crude mean in this population has a counterfactual interpretation.
- For example, suppose L is binary and we observe :

N	A	L	$E(Y\|A=a,L=l)$
4000	1	0	24
3000	1	1	36
8000	0	0	10
9000	0	1	22

- So that $f_{A|L}(A=1|L=1)=1/4, f_{A|L}(A=1|L=0)=1/3$ and thus L predicts A.
- Moreover $E(Y|A=1,L=l)=24+12l$ so that L predict Y given A.

G-computation:dual representation

- The crude mean
  $E\left(Y|A=1\right)=\sum_l E\left(Y|A=1,L=l\right)f\left(L=l|A=1\right)=$
  $24\times 4/7+36\times 3/7=204/7$

G-computation:dual representation

- The crude mean
  $E\left(Y|A=1\right)=\sum_l E\left(Y|A=1,L=l\right)f\left(L=l|A=1\right)=$
  $24\times 4/7+36\times 3/7=204/7$
- Whereas $E\left(Y_1\right)=\sum_l E\left(Y|A=1,L=l\right)f\left(L=l\right)=$
  $24\times 1/2+36\times 1/2=210/7$

G-computation:dual representation

- The crude mean
  $E\left(Y|A=1\right)=\sum_l E\left(Y|A=1,L=l\right)f\left(L=l|A=1\right)=$
  $24\times 4/7+36\times 3/7=204/7$
- Whereas $E\left(Y_1\right)=\sum_l E\left(Y|A=1,L=l\right)f\left(L=l\right)=$
  $24\times 1/2+36\times 1/2=210/7$
- Create a pseudo population by reweigthing the numbers in each row by $f_{A|L}^{-1}$

N	$f(A\|L)$	$Pseudo-N$	A	L	$E\left(Y\|A=a,L=l\right)$
4000	$1/3$	$4000\times 3=12,000$	1	0	24
3000	$1/4$	$3000\times 4=12,000$	1	1	36
8000	$2/3$	$8000\times 3/2=12,000$	0	0	10
9000	$3/4$	$9000\times 4/3=12,000$	0	1	22

G-computation:dual representation

- The crude mean
  $E\left(Y|A=1\right)=\sum_l E\left(Y|A=1,L=l\right)f\left(L=l|A=1\right)=$
  $24\times 4/7+36\times 3/7=204/7$
- Whereas $E\left(Y_1\right)=\sum_l E\left(Y|A=1,L=l\right)f\left(L=l\right)=$
  $24\times 1/2+36\times 1/2=210/7$
- Create a pseudo population by reweigthing the numbers in each row by $f^{-1}_{A|L}$

N	$f(A\|L)$	$Pseudo-N$	A	L	$E\left(Y\|A=a,L=l\right)$
4000	$1/3$	$4000\times 3=12,000$	1	0	24
3000	$1/4$	$3000\times 4=12,000$	1	1	36
8000	$2/3$	$8000\times 3/2=12,000$	0	0	10
9000	$3/4$	$9000\times 4/3=12,000$	0	1	22

- So that in the Pseudo population, the crude analysis gives:$E^*\left(Y|A=1\right)=24\times 1/2+36\times 1/2=210/7=E\left(Y_1\right)$

G-computation:dual representation

- The crude mean
  $E\left(Y|A=1\right)=\sum_l E\left(Y|A=1,L=l\right)f\left(L=l|A=1\right)=$
  $24\times 4/7+36\times 3/7=204/7$
- Whereas $E\left(Y_1\right)=\sum_l E\left(Y|A=1,L=l\right)f\left(L=l\right)=$
  $24\times 1/2+36\times 1/2=210/7$
- Create a pseudo population by reweigthing the numbers in each row by $f^{-1}_{A|L}$

N	$f(A\|L)$	$Pseudo-N$	A	L	$E\left(Y\|A=a,L=l\right)$
4000	$1/3$	$4000\times 3=12,000$	1	0	24
3000	$1/4$	$3000\times 4=12,000$	1	1	36
8000	$2/3$	$8000\times 3/2=12,000$	0	0	10
9000	$3/4$	$9000\times 4/3=12,000$	0	1	22

- So that in the Pseudo population, the crude analysis gives:$E^*\left(Y|A=1\right)=24\times 1/2+36\times 1/2=210/7=E\left(Y_1\right)$
- Do a similar calculation for $E\left(Y_0\right)$

Parametric G-computation.

- As before, the goal is to estimate the average causal effect:

$$\begin{aligned}\psi &= g(1) - g(0) \\ &= \int \{E(Y|A=1, L=l) - E(Y|A=a, L=l)\} \, dF_L(l)\end{aligned}$$

Parametric G-computation.

- As before, the goal is to estimate the average causal effect:

$$\begin{aligned}\psi &= g(1) - g(0) \\ &= \int \{E(Y|A=1, L=l) - E(Y|A=a, L=l)\} dF_L(l)\end{aligned}$$

- When L consists of a large number of categorical variables, most cells will contain few observations in moderate sample size data sets; so that the nonparametric estimator of $E(Y|A=a, L)$ will be unreliable .

Parametric G-computation.

- As before, the goal is to estimate the average causal effect:

$$\begin{aligned}\psi &= g(1) - g(0) \\ &= \int \{E(Y|A=1, L=l) - E(Y|A=a, L=l)\} dF_L(l)\end{aligned}$$

- When L consists of a large number of categorical variables, most cells will contain few observations in moderate sample size data sets; so that the nonparametric estimator of $E(Y|A=a, L)$ will be unreliable .
- Similarly, if L is a vector of continuous covariates,nonparametric smoothing is required to model $E(Y|A=a, L)$ nonparametrically, but because high dimensional multivariate smoothing convergence rates are inherently exceedingly slow, nonparametric regression will lead to unreliable estimates in moderate sample size.

Parametric G-computation.

- As before, the goal is to estimate the average causal effect:

$$\begin{aligned}\psi &= g(1) - g(0) \\ &= \int \{E(Y|A=1, L=l) - E(Y|A=a, L=l)\} dF_L(l)\end{aligned}$$

- When L consists of a large number of categorical variables, most cells will contain few observations in moderate sample size data sets; so that the nonparametric estimator of $E(Y|A=a, L)$ will be unreliable .
- Similarly, if L is a vector of continuous covariates,nonparametric smoothing is required to model $E(Y|A=a, L)$ nonparametrically, but because high dimensional multivariate smoothing convergence rates are inherently exceedingly slow, nonparametric regression will lead to unreliable estimates in moderate sample size.
- The debilitating effect of high dimensional L on nonparametric inference of ψ is commonly referred to as the 'curse of dimensionality'.

Parametric G-computation

- To circumvent this dimensionality problem, the statistician must make additional modelling assumptions .

Parametric G-computation

- To circumvent this dimensionality problem, the statistician must make additional modelling assumptions .
- One simple proposal is to posit a parametric model for $E(Y|A, L)$, say $b(A, L; \eta)$ indexed by a finite dimensional parameter η.

Parametric G-computation

- To circumvent this dimensionality problem, the statistician must make additional modelling assumptions .
- One simple proposal is to posit a parametric model for $E(Y|A, L)$, say $b(A, L; \eta)$ indexed by a finite dimensional parameter η.
- For instance, for continuous Y, the linear regression model $b(A, L; \eta) = (A, L')\,\eta$ is commonly used

Parametric G-computation

- To circumvent this dimensionality problem, the statistician must make additional modelling assumptions .
- One simple proposal is to posit a parametric model for $E(Y|A, L)$, say $b(A, L; \eta)$ indexed by a finite dimensional parameter η.
- For instance, for continuous Y, the linear regression model $b(A, L; \eta) = (A, L')\,\eta$ is commonly used
- For binary Y, one may use the logistic regression model $b(A, L; \eta) = (1 + \exp(-(A, L')\,\eta))^{-1}$ and obtain $\widehat{\eta}$ by maximizing the logistic likelihood function.

Parametric G-computation

- For a linear model $b(A, L; \eta)$,the standard approach to obtain estimates $\widehat{\eta}$ of η is with ordinary least-square which solve the normal equations

$$\sum_{i=1}^{n} (A_i, L_i') \left(Y_i - \left(A, L' \right) \eta \right) = 0$$

Parametric G-computation

- For a linear model $b(A, L; \eta)$,the standard approach to obtain estimates $\widehat{\eta}$ of η is with ordinary least-square which solve the normal equations

$$\sum_{i=1}^{n} (A_i, L_i') \left(Y_i - \left(A, L' \right) \eta \right) = 0$$

- Parametric G-computation then entails estimating ψ with

$$\widehat{\psi}_{or} = n^{-1} \sum_{i=1}^{n} \left(b(1, L_i; \widehat{\eta}) - b(0, L_i; \widehat{\eta}) \right)$$

Parametric G-computation

- For a linear model $b(A, L; \eta)$,the standard approach to obtain estimates $\widehat{\eta}$ of η is with ordinary least-square which solve the normal equations

$$\sum_{i=1}^{n} (A_i, L_i') \left(Y_i - \left(A, L' \right) \eta \right) = 0$$

- Parametric G-computation then entails estimating ψ with

$$\widehat{\psi}_{or} = n^{-1} \sum_{i=1}^{n} \left(b(1, L_i; \widehat{\eta}) - b(0, L_i; \widehat{\eta}) \right)$$

- In contrast to its nonparametric counterpart, the parametric G-computation estimator described above does not generally have a nice dual representation as an iptw estimator.

Parametric G-computation

- For a linear model $b(A, L; \eta)$,the standard approach to obtain estimates $\widehat{\eta}$ of η is with ordinary least-square which solve the normal equations

$$\sum_{i=1}^{n} (A_i, L_i') \left(Y_i - \left(A, L' \right) \eta \right) = 0$$

- Parametric G-computation then entails estimating ψ with

$$\widehat{\psi}_{or} = n^{-1} \sum_{i=1}^{n} \left(b(1, L_i; \widehat{\eta}) - b(0, L_i; \widehat{\eta}) \right)$$

- In contrast to its nonparametric counterpart, the parametric G-computation estimator described above does not generally have a nice dual representation as an iptw estimator.
- Moreover, it is susceptible to model misspecification as the regression model for $E(Y|A, L)$ will invariably be hard to correctly specify when L is high dimensional; for instance, if we have omitted higher order interactions or nonlinear regressors; therefore, a different approach is needed.

Parametric iptw estimation

- An alternative to parametric g-computation is to use the parametric iptw estimator with weights modelled using a parsimonuous parametric model;

Parametric iptw estimation

- An alternative to parametric g-computation is to use the parametric iptw estimator with weights modelled using a parsimonuous parametric model;
- That is, specify a model $\pi(L;\alpha)$ for the conditional probability of taking treatment $f_{A|L}(1|L)$, say $\pi(L;\alpha) = (1+\exp(-(L')\alpha))^{-1}$ indexed by a finite dimensional vector α. $\pi(L;\alpha)$ is a standard logistic regression thus $\widehat{\alpha}$ may be obtained from maximum likelihood theory with data (A_i, L_i) $i = 1, ..., n$.

Parametric iptw estimation

- The parametric iptw estimator of ψ is defined as

$$\widehat{\psi}_{iptw} = \frac{\sum_{s=1}^{n} I(A_s = 1) Y_s \pi^{-1}(L_s; \widehat{\alpha})}{\sum_{s=1}^{n} I(A_s = 1) \pi^{-1}(L_s; \widehat{\alpha})} - \frac{\sum_{s=1}^{n} I(A_s = 0) Y_s (1 - \pi(L_s; \widehat{\alpha}))^{-1}}{\sum_{s=1}^{n} I(A_s = 0) (1 - \pi(L_s; \widehat{\alpha}))^{-1}}$$

Parametric iptw estimation

- The parametric iptw estimator of ψ is defined as

$$\widehat{\psi}_{iptw} = \frac{\sum_{s=1}^{n} I(A_s = 1) Y_s \pi^{-1}(L_s; \widehat{\alpha})}{\sum_{s=1}^{n} I(A_s = 1) \pi^{-1}(L_s; \widehat{\alpha})} - \frac{\sum_{s=1}^{n} I(A_s = 0) Y_s (1 - \pi(L_s; \widehat{\alpha}))^{-1}}{\sum_{s=1}^{n} I(A_s = 0) (1 - \pi(L_s; \widehat{\alpha}))^{-1}}$$

- Whenever L is high dimensional, it will generally be difficult to correctly specify a parsimonuous model for $f_{A|L}(1|L)$, therefore $\widehat{\psi}$ may end up being significantly biased due to model misspecification.

Doubly robust estimation

- We have described two alternate ways of modelling the average treatment effect when the dimension of L is prohibitively large to allow for nonparametric methods; the methods described are the parametric g-computation and the parametric iptw.

Doubly robust estimation

- We have described two alternate ways of modelling the average treatment effect when the dimension of L is prohibitively large to allow for nonparametric methods; the methods described are the parametric g-computation and the parametric iptw.
- A question arises as to which of the two estimators is to be prefered for any given data set, since they make different modelling assumptions.

Doubly robust estimation

- We have described two alternate ways of modelling the average treatment effect when the dimension of L is prohibitively large to allow for nonparametric methods; the methods described are the parametric g-computation and the parametric iptw.
- A question arises as to which of the two estimators is to be prefered for any given data set, since they make different modelling assumptions.
- Because we can never be certain that we have adequately modelled either the outcome regression (or) or the treatment process (also known as the propensity score), the best we can hope for, is to find an estimator that is consistent and asymptotically normal when either but not necessarily both parametric models are correctly specified.

Doubly robust estimation

- We have described two alternate ways of modelling the average treatment effect when the dimension of L is prohibitively large to allow for nonparametric methods; the methods described are the parametric g-computation and the parametric iptw.
- A question arises as to which of the two estimators is to be prefered for any given data set, since they make different modelling assumptions.
- Because we can never be certain that we have adequately modelled either the outcome regression (or) or the treatment process (also known as the propensity score), the best we can hope for, is to find an estimator that is consistent and asymptotically normal when either but not necessarily both parametric models are correctly specified.
- An estimator that satisfies this latter property is known as a *doubly robust* estimator.

Doubly robust estimation

- Given estimated parametric models $\pi\left(L;\widehat{\alpha}\right)$, $b(A,L;\widehat{\eta})$, a doubly robust estimator of the average causal effect is given by :

$$\begin{aligned}\widehat{\psi}_{dr} &= n^{-1}\sum_i\left(b(1,L_i;\widehat{\eta})+\frac{I\left(A_i=1\right)}{\pi\left(L_i;\widehat{\alpha}\right)}\left(Y_i-b(A_i,L_i;\widehat{\eta})\right)\right)\\ &\quad -n^{-1}\sum_i\left(b(0,L_i;\widehat{\eta})+\frac{I\left(A_i=0\right)}{\left(1-\pi\left(L_i;\widehat{\alpha}\right)\right)}\left(Y_i-b(A_i,L_i;\widehat{\eta})\right)\right)\end{aligned}$$

Doubly robust estimation

- Given estimated parametric models $\pi\left(L;\widehat{\alpha}\right)$, $b(A,L;\widehat{\eta})$, a doubly robust estimator of the average causal effect is given by :

$$\begin{aligned}\widehat{\psi}_{dr} &= n^{-1}\sum_i\left(b(1,L_i;\widehat{\eta})+\frac{I\left(A_i=1\right)}{\pi\left(L_i;\widehat{\alpha}\right)}\left(Y_i-b(A_i,L_i;\widehat{\eta})\right)\right)\\ &\quad -n^{-1}\sum_i\left(b(0,L_i;\widehat{\eta})+\frac{I\left(A_i=0\right)}{\left(1-\pi\left(L_i;\widehat{\alpha}\right)\right)}\left(Y_i-b(A_i,L_i;\widehat{\eta})\right)\right)\end{aligned}$$

- One can show that $\widehat{\psi}_{dr}$ is consistent if either $E(Y|A,L)=(A,L')\,\eta$, or $f_{A|L}\left(A|L_i\right)=\pi\left(L;\alpha\right)$ but not necessarily both are correctly specified.

Part II:Longitudinal Causal Effects

Causal Diagrams

- Consider a hypothetical study of the relation of antihistamine treatment to asthma incidence among first-grade children among first-grade public-school children (From Greenland et al. 1999)

Causal Diagrams

- Consider a hypothetical study of the relation of antihistamine treatment to asthma incidence among first-grade children among first-grade public-school children (From Greenland et al. 1999)
- Let A=air pollution level,B=sex, C=Bronchial reactivity,D=asthma, E=antihistamine.

Causal Diagrams

- Consider a hypothetical study of the relation of antihistamine treatment to asthma incidence among first-grade children among first-grade public-school children (From Greenland et al. 1999)
- Let A=air pollution level,B=sex, C=Bronchial reactivity,D=asthma, E=antihistamine.
- Suppose we further know that :

Causal Diagrams

- Consider a hypothetical study of the relation of antihistamine treatment to asthma incidence among first-grade children among first-grade public-school children (From Greenland et al. 1999)
- Let A=air pollution level,B=sex, C=Bronchial reactivity,D=asthma, E=antihistamine.
- Suppose we further know that :
  1. pollution is independent of sex

Causal Diagrams

- Consider a hypothetical study of the relation of antihistamine treatment to asthma incidence among first-grade children among first-grade public-school children (From Greenland et al. 1999)
- Let A=air pollution level,B=sex, C=Bronchial reactivity,D=asthma, E=antihistamine.
- Suppose we further know that :
  1. pollution is independent of sex
  2. sex affects administration of antihistamine only through bronchial reactivity, but directly influences asthma risk

Causal Diagrams

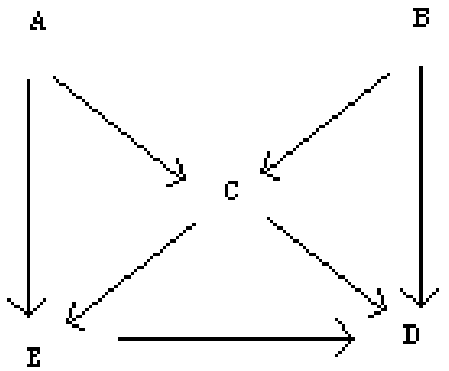
- Consider a hypothetical study of the relation of antihistamine treatment to asthma incidence among first-grade children among first-grade public-school children (From Greenland et al. 1999)
- Let A=air pollution level,B=sex, C=Bronchial reactivity,D=asthma, E=antihistamine.
- Suppose we further know that :
  1. pollution is independent of sex
  2. sex affects administration of antihistamine only through bronchial reactivity, but directly influences asthma risk
  3. Industrial air pollution only leads to asthma attacks through antihistamine use and bronchial reactivity.

Causal Diagrams

- Consider a hypothetical study of the relation of antihistamine treatment to asthma incidence among first-grade children among first-grade public-school children (From Greenland et al. 1999)
- Let A=air pollution level,B=sex, C=Bronchial reactivity,D=asthma, E=antihistamine.
- Suppose we further know that :
  1. pollution is independent of sex
  2. sex affects administration of antihistamine only through bronchial reactivity, but directly influences asthma risk
  3. Industrial air pollution only leads to asthma attacks through antihistamine use and bronchial reactivity.
  4. And there are no other important confounders besides sex, bronchial reactivity and air pollution

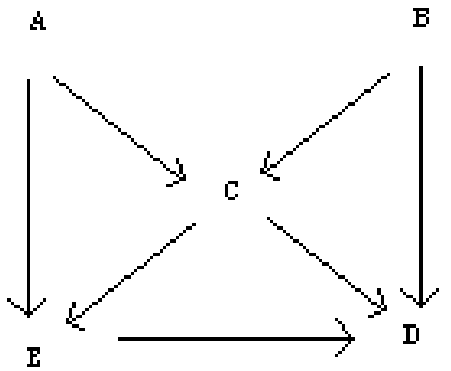
Causal Diagrams

- These assertions can be incorporated in the following diagram



Causal Diagrams

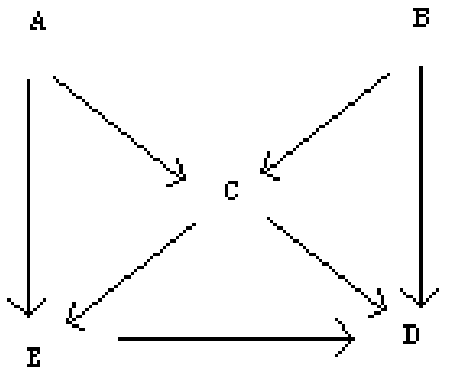
- These assertions can be incorporated in the following diagram



  - The points representing the variables are called the vertices/nodes of the graph; any line or arrow connecting two variables in the graph is an arc/edge.

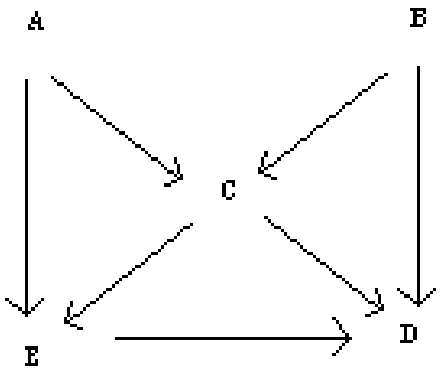
Causal Diagrams

- These assertions can be incorporated in the following diagram



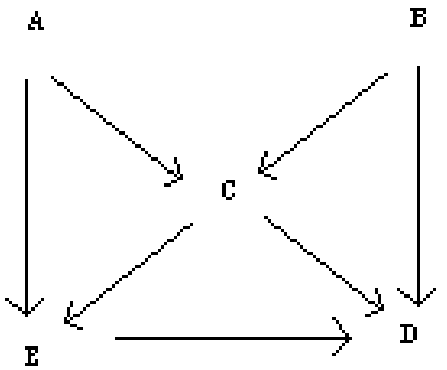
  - The points representing the variables are called the vertices/nodes of the graph; any line or arrow connecting two variables in the graph is an arc/edge.
  - Arrows represent direct links from causes to effects, that is not mediated by any other variable. Example: the arrow linking A and C represents a direct effect of A on C.

Causal Diagrams



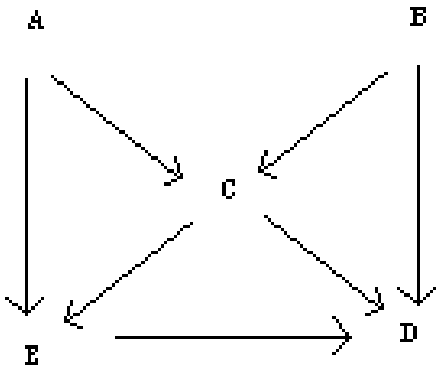
- Absence of arrow $\Leftrightarrow$no direct causal effect.Example:. no arrow from A to D reflects assertion (3) above.

Causal Diagrams



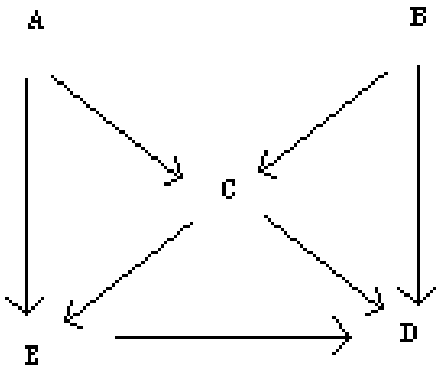
- Absence of arrow $\Leftrightarrow$no direct causal effect.Example:. no arrow from A to D reflects assertion (3) above.
- a node within a path is said to *intercept* the path: Example C intercepts the paths A-C-D and E-C-D.

Causal Diagrams



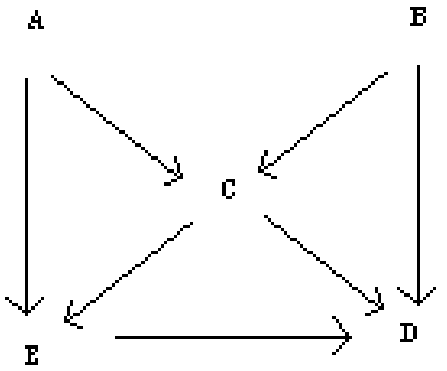
- X is an *ancestor or cause* of Y$\Leftrightarrow$there is a directed path leading out of X into Y. So that Y is a descendant of X. Example: A.B and C are ancestors of E and D, which in turn are descendants of A,B and C.

Causal Diagrams



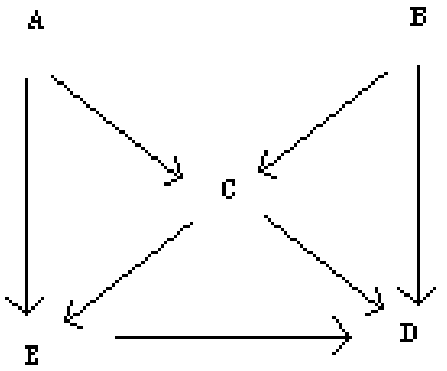
- X is an *ancestor or cause* of Y$\Leftrightarrow$there is a directed path leading out of X into Y. So that Y is a descendant of X. Example: A.B and C are ancestors of E and D, which in turn are descendants of A,B and C.
- X is a parent of Y if there is a single headed arrow from X into Y: in such a case Y is called a child of X. Example: A and C are parents of E, whereas C and E are children of A.

Causal Diagrams



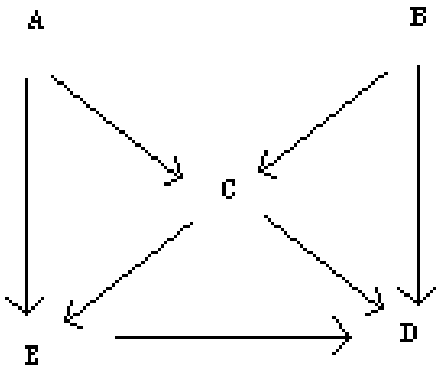
- A path that connects X to Y is a *backdoor path* from X to Y if it has an arrowhead pointing to X. Example: all paths from E to D except the direct path.

Causal Diagrams



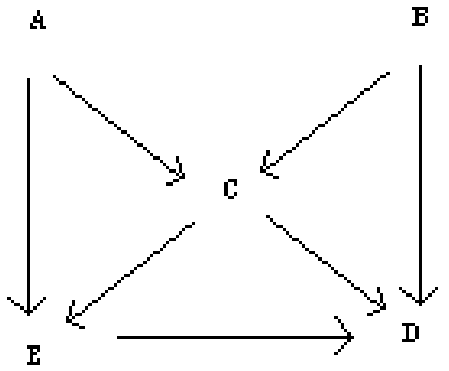
- A path that connects X to Y is a *backdoor path* from X to Y if it has an arrowhead pointing to X. Example: all paths from E to D except the direct path.
- A path *collides* at a variable X if the path enters and exists X through arrowheads in which case X is called a *collider* on the path. A path is *blocked* if it has one or more colliders, otherwise it is *unblocked.* Examples:

Causal Diagrams



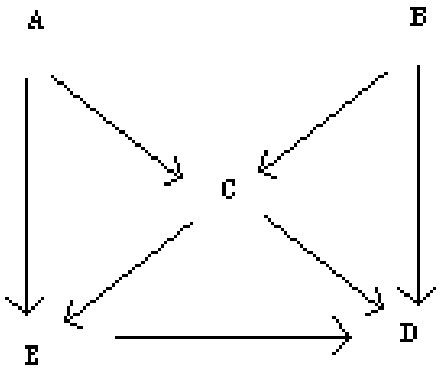
- A path that connects X to Y is a *backdoor path* from X to Y if it has an arrowhead pointing to X. Example: all paths from E to D except the direct path.
- A path *collides* at a variable X if the path enters and exists X through arrowheads in which case X is called a *collider* on the path. A path is *blocked* if it has one or more colliders, otherwise it is *unblocked.* Examples:
  - The backdoor path E-A-C-B-D is blocked because it collides at C

Causal Diagrams



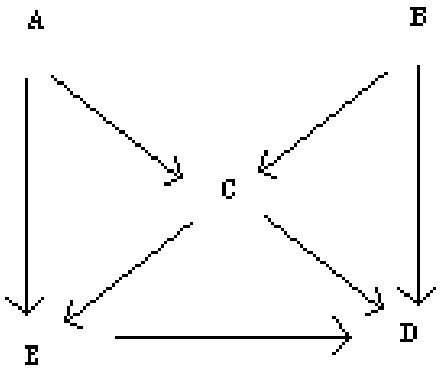
- A path that connects X to Y is a *backdoor path* from X to Y if it has an arrowhead pointing to X. Example: all paths from E to D except the direct path.
- A path *collides* at a variable X if the path enters and exists X through arrowheads in which case X is called a *collider* on the path. A path is *blocked* if it has one or more colliders, otherwise it is *unblocked.* Examples:
  - The backdoor path E-A-C-B-D is blocked because it collides at C
  - the backdoor path E-A-C-D is unblocked because it contains no collider.

Causal Diagrams



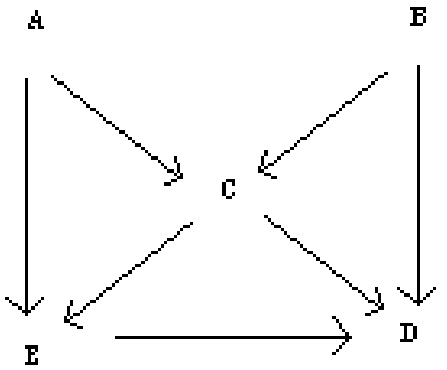
- The diagram for this study is a *Directed Acyclic Graph* (DAG)

Causal Diagrams



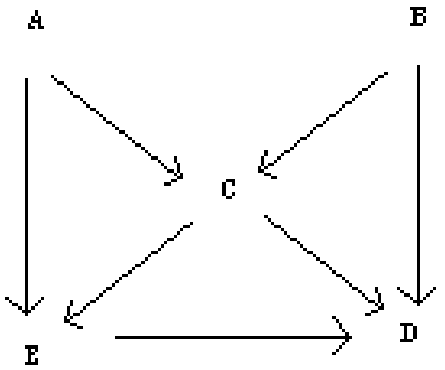
- The diagram for this study is a *Directed Acyclic Graph* (DAG)
  - *Directed* since all arcs between variable are arrows.Moreover directed path$\Leftrightarrow$causal path.

Causal Diagrams



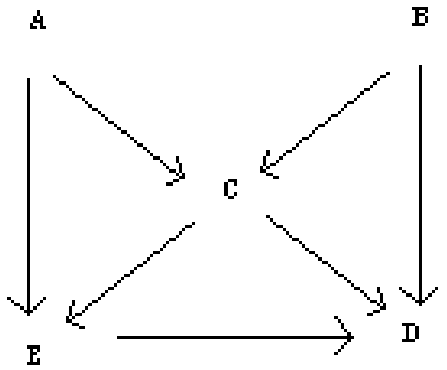
- The diagram for this study is a *Directed Acyclic Graph* (DAG)
  - *Directed* since all arcs between variable are arrows.Moreover directed path⇔causal path.
  - *Acyclic* if no directed path (enter through the tail, leave throught the head) in the graph foms a closed loop.

Causal Diagrams



- The diagram for this study is a *Directed Acyclic Graph* (DAG)
  - *Directed* since all arcs between variable are arrows.Moreover directed path⇔causal path.
  - *Acyclic* if no directed path (enter through the tail, leave throught the head) in the graph foms a closed loop.
- A DAG is a causal DAG if all common causes of any pair of variables in the graph are also in the graph.

Causal Diagrams



- The diagram for this study is a *Directed Acyclic Graph* (DAG)
  - *Directed* since all arcs between variable are arrows.Moreover directed path⇔causal path.
  - *Acyclic* if no directed path (enter through the tail, leave throught the head) in the graph foms a closed loop.
- A DAG is a causal DAG if all common causes of any pair of variables in the graph are also in the graph.
  - does not need to include variables not of interest or not common causes of variables in the DAG.

Causal Diagrams

- A DAG is nonparametric$\Leftrightarrow$does not impose any functional restriction on the joint distribution of the variables on the graph.

Causal Diagrams

- A DAG is nonparametric$\Leftrightarrow$does not impose any functional restriction on the joint distribution of the variables on the graph.
- Production of an effect by a cause requires a directed path from the cause to the effect on the graph.e.g. Absence of a directed path between A and B implies absence of a causal effect between the two variables, which also implies there's no effect of A on D through B.

Causal Diagrams

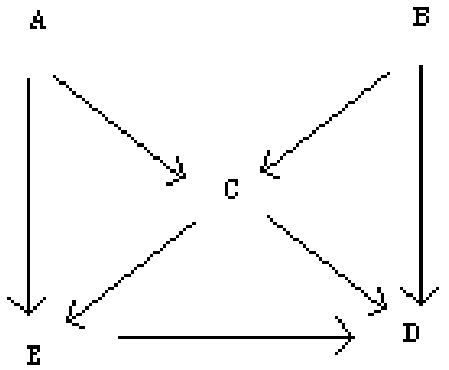
- A DAG is nonparametric$\Leftrightarrow$does not impose any functional restriction on the joint distribution of the variables on the graph.
- Production of an effect by a cause requires a directed path from the cause to the effect on the graph.e.g. Absence of a directed path between A and B implies absence of a causal effect between the two variables, which also implies there's no effect of A on D through B.
- Graphs also encode 'associations' between variables: absence of an unblocked path between two variables $\Rightarrow$statistical independence variables. e.g: A and B are marginally independent. In other words,marginally associated covariates require the presence of an unblocked path on the graph.

Causal Diagrams

- Given a causal DAG, we can deduce implied conditional independences in the observed data:

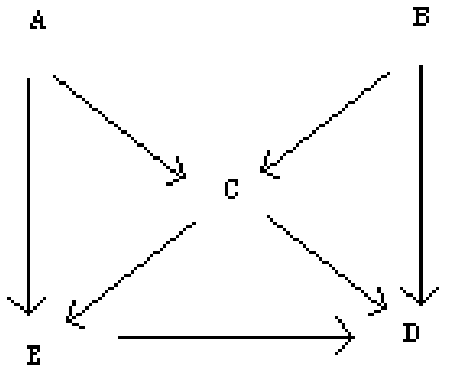
Causal Diagrams

- Given a causal DAG, we can deduce implied conditional independences in the observed data:
- For instance in our causal DAG,



Causal Diagrams

- Given a causal DAG, we can deduce implied conditional independences in the observed data:
- For instance in our causal DAG,

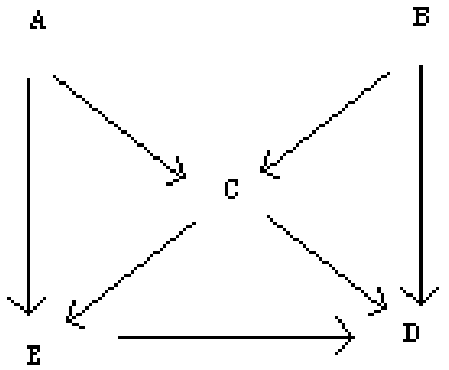


- Observed data pdf factorizes as:

$$f(A, E, C, B, D) = f(D|B, C, E)\, f(C|A, B)\, f(E|A, C) f(A)\, f(B)$$

Causal Diagrams

- Given a causal DAG, we can deduce implied conditional independences in the observed data:
- For instance in our causal DAG,



- Observed data pdf factorizes as:

$$f(A, E, C, B, D) = f(D|B, C, E)\, f(C|A, B)\, f(E|A, C) f(A)\, f(B)$$

- Where we use the markov factorization, that a variable is independent of nonparental ancestors given its parents:

$$f(A, E, C, B, D) = f(D|pa(D))\, f(C|pa(C))\, f(E|pa(E)) f(A|pa(A))$$

Causal Diagrams

- In a DAG, only two kinds of unblocked paths can occur:

Causal Diagrams

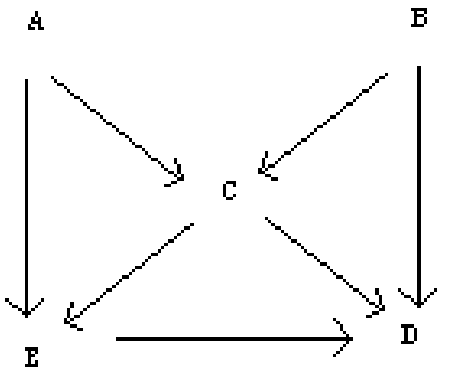
- In a DAG, only two kinds of unblocked paths can occur:
  - a directed path: which holds if association is at least partly causal, thus the effect is descendant of the cause.

Causal Diagrams

- In a DAG, only two kinds of unblocked paths can occur:
  - a directed path: which holds if association is at least partly causal, thus the effect is descendant of the cause.
  - a backdoor path through a shared ancestor: which holds if association is at least partly confounded.

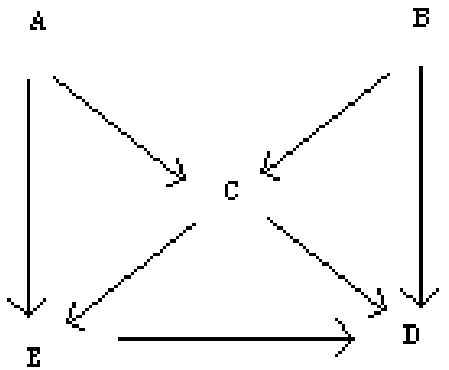
Causal Diagrams

- In a DAG, only two kinds of unblocked paths can occur:
  - a directed path: which holds if association is at least partly causal, thus the effect is descendant of the cause.
  - a backdoor path through a shared ancestor: which holds if association is at least partly confounded.
- Of course, both conditions may hold as with E and D.



Causal Diagrams

- In a DAG, only two kinds of unblocked paths can occur:
  - a directed path: which holds if association is at least partly causal, thus the effect is descendant of the cause.
  - a backdoor path through a shared ancestor: which holds if association is at least partly confounded.
- Of course, both conditions may hold as with E and D.



- Note that E-A-C-B-D is blocked at the collider C, but E-A-C-D and E-C-B-D are both unblocked backdoor paths.

Causal Diagrams

- Note also that the presence of an unblocked path between two variables is meant to allow but does not necessarily imply an association between them.

Causal Diagrams

- Note also that the presence of an unblocked path between two variables is meant to allow but does not necessarily imply an association between them.
- For instance, the three backdoor paths between E and D could cancel out with the direct path to yield no marginal association between E and D.

Causal Diagrams

- Note also that the presence of an unblocked path between two variables is meant to allow but does not necessarily imply an association between them.
- For instance, the three backdoor paths between E and D could cancel out with the direct path to yield no marginal association between E and D.
- It should be clear that the presence or absence of blocked paths should not affect the association between variables. This is because the marginal association between two causes of an effect (ancestors of a collider) is fixed by the time both causes have occured; i.e. this association cannot be affected by consequences of these variables.

DAGS and Confounding

- Working definition: *Confounding* occurs when the study exposure groups differ in their probability distribution for the outcome for reasons other than exposure effect.

DAGS and Confounding

- Working definition: *Confounding* occurs when the study exposure groups differ in their probability distribution for the outcome for reasons other than exposure effect.
- Such differences are attributable to effects of extraneous variables which are called *confounders*.

DAGS and Confounding

- Working definition: *Confounding* occurs when the study exposure groups differ in their probability distribution for the outcome for reasons other than exposure effect.
- Such differences are attributable to effects of extraneous variables which are called *confounders*.
- Counfounding is present if and only if exposure would remain associated with disease even if all exposure effects were removed, prevented or blocked.

DAGS and Confounding

- This condition is easy to check in a DAG that represents relations among exposure,disease and potential confounders with the following algorithm

DAGS and Confounding

- This condition is easy to check in a DAG that represents relations among exposure,disease and potential confounders with the following algorithm
  1. Delete all exposure effects.

DAGS and Confounding

- This condition is easy to check in a DAG that represents relations among exposure,disease and potential confounders with the following algorithm
  1. Delete all exposure effects.
  2. In the new graph without exposure effects,check whether there is any unblocked path from exposure to disease.

DAGS and Confounding

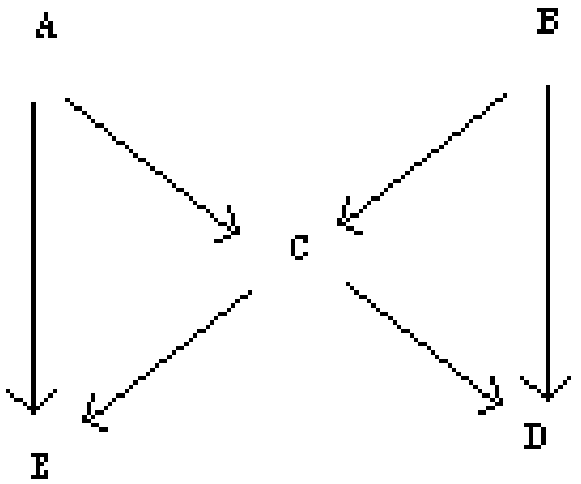
- This condition is easy to check in a DAG that represents relations among exposure,disease and potential confounders with the following algorithm
  1. Delete all exposure effects.
  2. In the new graph without exposure effects,check whether there is any unblocked path from exposure to disease.
- This algorithm checks whether exposure and disease would remain associated under the null of no causal effect of E on D, i.e. do they share a common ancestor?

DAGS and Confounding

- This condition is easy to check in a DAG that represents relations among exposure,disease and potential confounders with the following algorithm
  1. Delete all exposure effects.
  2. In the new graph without exposure effects,check whether there is any unblocked path from exposure to disease.
- This algorithm checks whether exposure and disease would remain associated under the null of no causal effect of E on D, i.e. do they share a common ancestor?
- Note that the effects of disease play no role in the above algorithm, since all paths from exposure to disease through descendants of disease must pass through a collider and are therefore blocked.

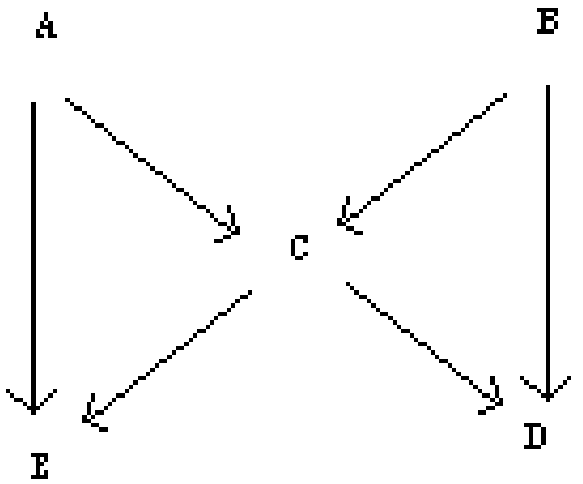
DAGS and Confounding

- Applying to our DAG, we see that A, C, and B are *potential confounders*



DAGS and Confounding

- Applying to our DAG, we see that A, C, and B are *potential confounders*



- The next natural question is whether and how one can control for confounding in assessing the effect of E on D.

DAGS and Confounding

- First lets review a conventional statistical approach for assessing confounding:

DAGS and Confounding

- First lets review a conventional statistical approach for assessing confounding:
  1. The variable is an ancestor of the cause

DAGS and Confounding

- First lets review a conventional statistical approach for assessing confounding:
  1. The variable is an ancestor of the cause
  2. The variable is associated with exposure

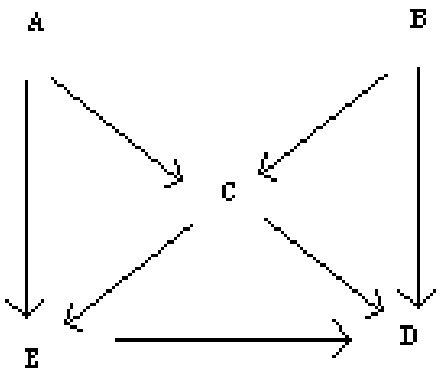
DAGS and Confounding

- First lets review a conventional statistical approach for assessing confounding:
  1. The variable is an ancestor of the cause
  2. The variable is associated with exposure
  3. The variable is not a descendant of the exposure or outcome.

DAGS and Confounding

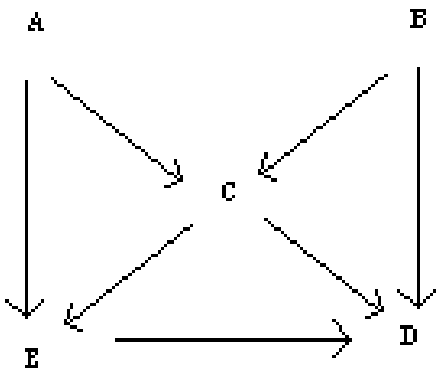
- First lets review a conventional statistical approach for assessing confounding:
  1. The variable is an ancestor of the cause
  2. The variable is associated with exposure
  3. The variable is not a descendant of the exposure or outcome.
- With these criteria in mind, what is the smallest subset of variables from A,B and C that would be sufficient to control for confounding?

DAGS and Confounding



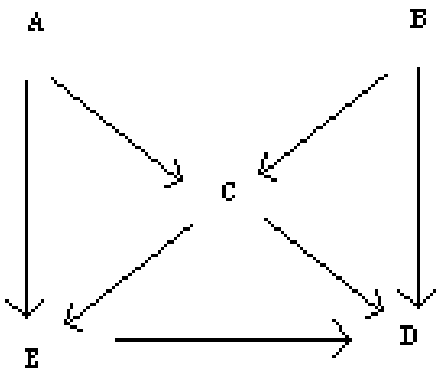
- A conventional approach is to condition on potential confonders:

DAGS and Confounding



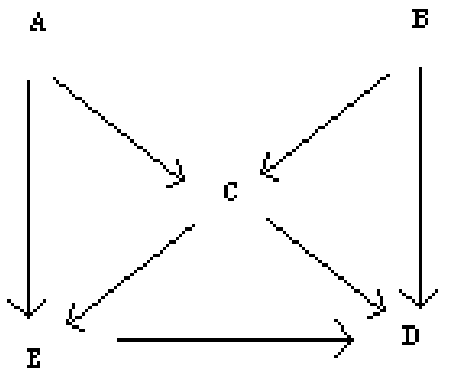
- A conventional approach is to condition on potential confonders:
  - Consider conditioning on A;clearly this blocks the path E-A-C-D, but E-C-D and E-C-B-D still unblocked.

DAGS and Confounding



- A conventional approach is to condition on potential confonders:
  - Consider conditioning on A;clearly this blocks the path E-A-C-D, but E-C-D and E-C-B-D still unblocked.
  - Similarly, conditioning on B blocks E-C-B-A, but E-C-D and E-A-C-D unblocked.

DAGS and Confounding



- A conventional approach is to condition on potential confonders:
  - Consider conditioning on A;clearly this blocks the path E-A-C-D, but E-C-D and E-C-B-D still unblocked.
  - Similarly, conditioning on B blocks E-C-B-A, but E-C-D and E-A-C-D unblocked.
  - Finally, conditioning on C alone seems promising as it blocks the path E-A-C-D, as well as the paths E-C-B-D and E-C-D. Thus standard logic would go as follows "... once we adjust for C, variables A and B would fail to satisfy one of the necessary conditions 2 and 3 required of confounders and therefore adjustment for C would control confounding by A and B as well as C".

DAGS and Confounding

- Is this right? Consider the following numerical example

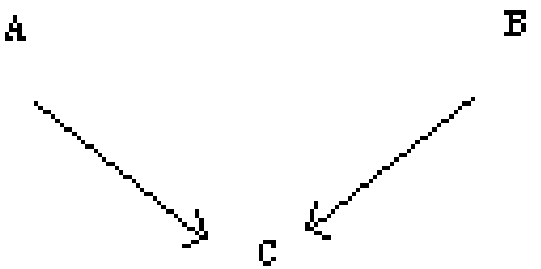
	$A = 1$		$A = 0$	
	$B = 1$	$B = 0$	$B = 1$	$B = 0$
$C = 1$	800	600	400	200
$C = 0$	200	400	600	800
Total	1000	1000	1000	1000

DAGS and Confounding

- Is this right? Consider the following numerical example

	$A=1$		$A=0$	
	$B=1$	$B=0$	$B=1$	$B=0$
$C=1$	800	600	400	200
$C=0$	200	400	600	800
Total	1000	1000	1000	1000

- We have
  $\Pr(A=1|B) = \Pr(A=1) = P(B=1|A) = P(B=1) = 0.5$. A, and B are marginally independent. Moreover, $P(C=1|A=1,B) - P(C=1|A=0,B) = .4$, and
  $P(C=1|A,B=1) - P(C=1|A,B=0) = .2$; which is consistent with the DAG



DAGS and Confounding

- Verify that the conditional odds ratio for the A-B association is 2/3 within strata of C.
  So that conditioning on C induces an association between A and B though they were marginally independent.

DAGS and Confounding

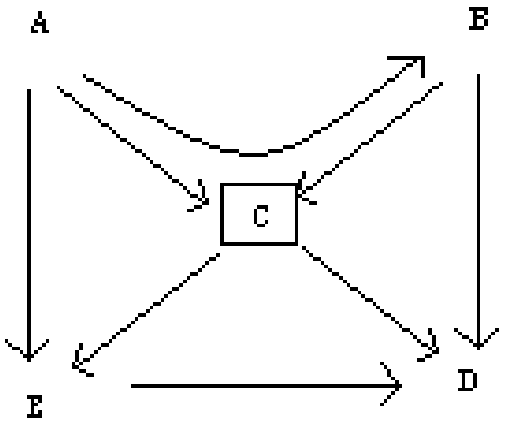
- Verify that the conditional odds ratio for the A-B association is 2/3 within strata of C.
  So that conditioning on C induces an association between A and B though they were marginally independent.
- This is an example of a general rule: If C is a common effect of A and B, then the association of A and B within levels of C will generally differ from the marginal association. This is a generalization of *Berkson's bias.*

DAGS and Confounding

- Returning to our original DAG, applying this rule tells us that conditioning on C can create an unblocked back door path from E to D: the association of A and B within levels of C can create an association of A and D indirectly, through B.

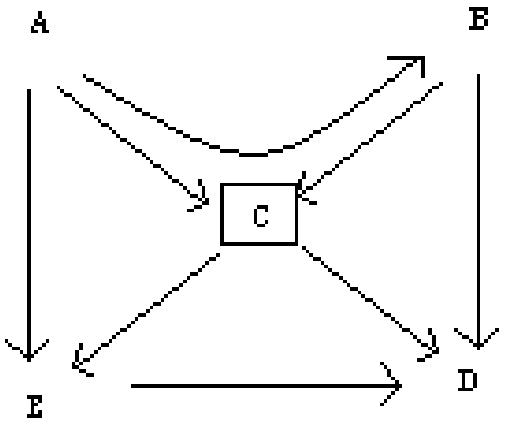
DAGS and Confounding

- Returning to our original DAG, applying this rule tells us that conditioning on C can create an unblocked back door path from E to D: the association of A and B within levels of C can create an association of A and D indirectly, through B.
- We conclude that it is not sufficient to only condition on either A, B or C to control for confounding.



DAGS and Confounding

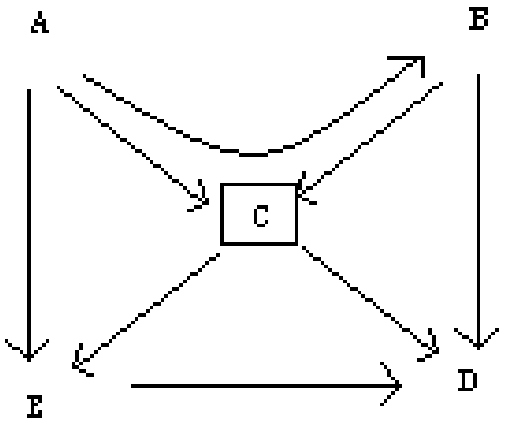
- Returning to our original DAG, applying this rule tells us that conditioning on C can create an unblocked back door path from E to D: the association of A and B within levels of C can create an association of A and D indirectly, through B.
- We conclude that it is not sufficient to only condition on either A, B or C to control for confounding.



- However, conditioning on either A and C or B and C is sufficient. Why???

DAGS and Confounding

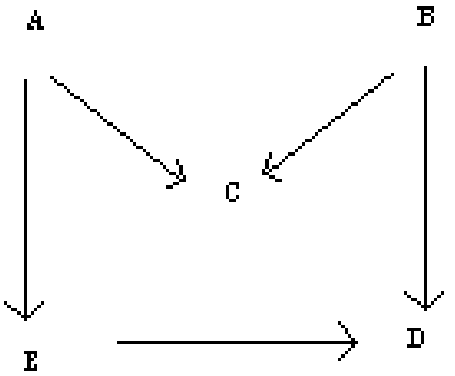
- Returning to our original DAG, applying this rule tells us that conditioning on C can create an unblocked back door path from E to D: the association of A and B within levels of C can create an association of A and D indirectly, through B.
- We conclude that it is not sufficient to only condition on either A, B or C to control for confounding.



- However, conditioning on either A and C or B and C is sufficient. Why???
- Moreover, this illustrates that the popular criteria used to assess confounders is necessary but not sufficient.

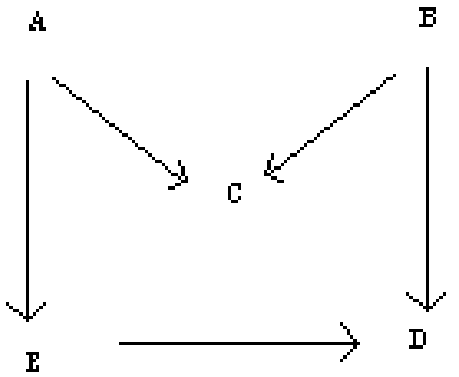
DAGS and Confounding

- Another example where things can go terribly wrong:



DAGS and Confounding

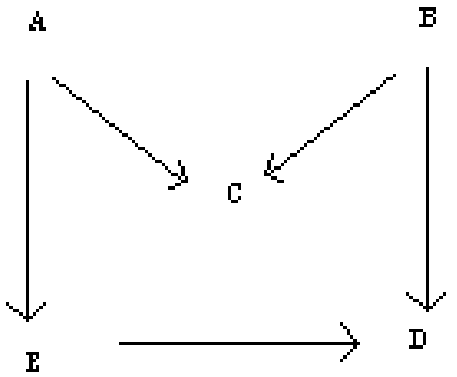
- Another example where things can go terribly wrong:



- Clearly there's no confounding according to the causal DAG given above;as the only backdoor path E-A-C-B-D is blocked by C. However, conventional wisdom will find that C is associated with E and associated with D given E, making it a potential confounder to adjust.

DAGS and Confounding

- Another example where things can go terribly wrong:



- Clearly there's no confounding according to the causal DAG given above;as the only backdoor path E-A-C-B-D is blocked by C. However, conventional wisdom will find that C is associated with E and associated with D given E, making it a potential confounder to adjust.
- In this DAG, conditioning on the collider C leads to confounding; worse if neither A nor B are observed as this would lead to untractable confounding.

DAGS and Confounding

- Take home message: statistical criteria are insufficient to characterize confounding, we need to first write down our causal DAG (explicit prior belief), from which we can decide which if any variables need to be conditioned on to control for confounding by using an "appropriate set of graphical rules".

DAGS and Confounding

- Take home message: statistical criteria are insufficient to characterize confounding, we need to first write down our causal DAG (explicit prior belief), from which we can decide which if any variables need to be conditioned on to control for confounding by using an "appropriate set of graphical rules".
- *D-separation* is one such rules:

DAGS and Confounding

- Take home message: statistical criteria are insufficient to characterize confounding, we need to first write down our causal DAG (explicit prior belief), from which we can decide which if any variables need to be conditioned on to control for confounding by using an "appropriate set of graphical rules".
- *D-separation* is one such rules:
  - It is a rule to decide whether two variables are either *d-separated* (independent) , or *d-connected* (associated)

DAGS and Confounding

- Take home message: statistical criteria are insufficient to characterize confounding, we need to first write down our causal DAG (explicit prior belief), from which we can decide which if any variables need to be conditioned on to control for confounding by using an "appropriate set of graphical rules".
- *D-separation* is one such rules:
  - It is a rule to decide whether two variables are either *d-separated* (independent) , or *d-connected* (associated)
  - If two variables are d-separated without conditioning on other variables, then they are marginally independent; if they are d-separated only after conditioning on a set of third variables S, they are conditionally independent given S.

DAGS and Confounding

- D-separation: We say that S *d-separates* two sets of variables R and T if the following hold:

DAGS and Confounding

- D-separation: We say that S *d-separates* two sets of variables R and T if the following hold:
  - Every unblocked path from R to T is intercepted by a variable in S.

DAGS and Confounding

- D-separation: We say that S *d-separates* two sets of variables R and T if the following hold:
  - Every unblocked path from R to T is intercepted by a variable in S.
  - Every unblocked path from R to T generated by adjustment for the variables in S is intercepted by a variable S.

DAGS and Confounding

- D-separation: We say that S *d-separates* two sets of variables R and T if the following hold:
  - Every unblocked path from R to T is intercepted by a variable in S.
  - Every unblocked path from R to T generated by adjustment for the variables in S is intercepted by a variable S.
- The *back-door* criteria:

DAGS and Confounding

- D-separation: We say that S *d-separates* two sets of variables R and T if the following hold:
  - Every unblocked path from R to T is intercepted by a variable in S.
  - Every unblocked path from R to T generated by adjustment for the variables in S is intercepted by a variable S.
- The *back-door* criteria:
  - a set of variables S is sufficient for control of confounding under a given DAG if S contains no descendants of E or D, and S d-separates E from D in the graph obtained by deleting all arrows emanating from E.

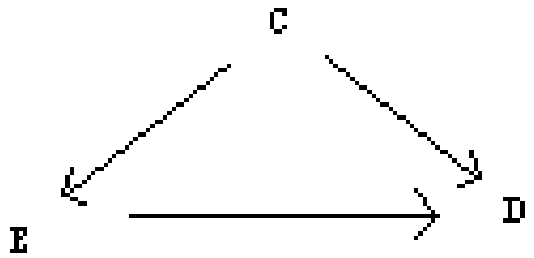
DAGS and Confounding

- D-separation: We say that S *d-separates* two sets of variables R and T if the following hold:
  - Every unblocked path from R to T is intercepted by a variable in S.
  - Every unblocked path from R to T generated by adjustment for the variables in S is intercepted by a variable S.
- The *back-door* criteria:
  - a set of variables S is sufficient for control of confounding under a given DAG if S contains no descendants of E or D, and S d-separates E from D in the graph obtained by deleting all arrows emanating from E.
  - Does this work? Certainly in the asthma study, back door criterion gives S={A,C} or S={B,C}.

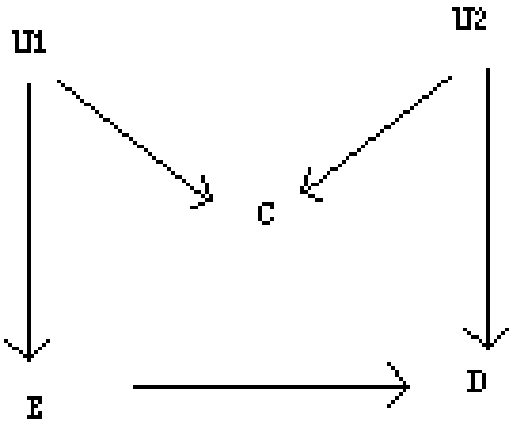
DAGS and Confounding

- Lets consider the following DAGs.

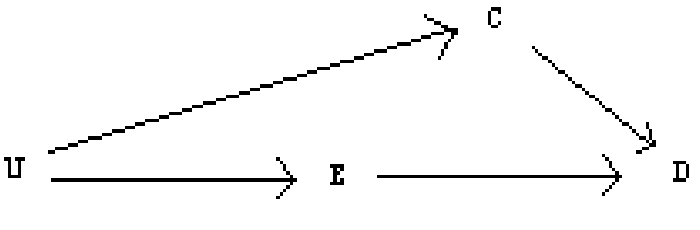
1)



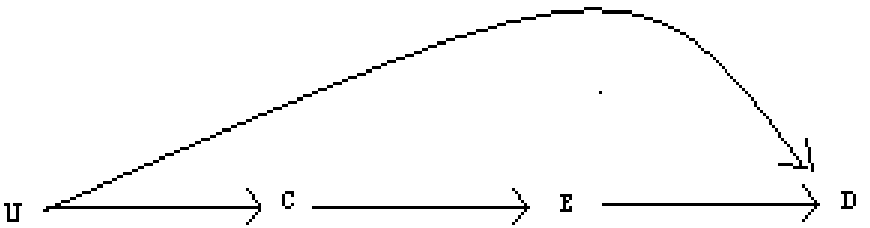
2)



3)



4)



DAGS and Confounding

- In Causal DAGS 1),2),3) and 4), give the corresponding markov factorization of their pdf.

DAGS and Confounding

- In Causal DAGS 1),2),3) and 4), give the corresponding markov factorization of their pdf.
- List the conditional independences implied by each causal DAG.

DAGS and Confounding

- In Causal DAGS 1),2),3) and 4), give the corresponding markov factorization of their pdf.
- List the conditional independences implied by each causal DAG.
- Suppose U variables were not collected by the investigator, using the backdoor criterion, is the effect of E on D identifiable in all DAGs?

Causal Inference for Longitudinal Data

- The Multicenter Aids Cohort Study (MACS) followed HIV+ men in 4 US cities,

Causal Inference for Longitudinal Data

- The Multicenter Aids Cohort Study (MACS) followed HIV+ men in 4 US cities,
- We restict our attention to data from1996 to 2002,that is the HAART era

Causal Inference for Longitudinal Data

- The Multicenter Aids Cohort Study (MACS) followed HIV+ men in 4 US cities,
- We restict our attention to data from1996 to 2002,that is the HAART era
- Data includes the following covariates measured at baseline and at every bi-annual visit:HAART use, CD4 cell count, plasma HIV-1 viral load,use of mono-therapy or combination antiretroviral therapy.

Causal Inference for Longitudinal Data

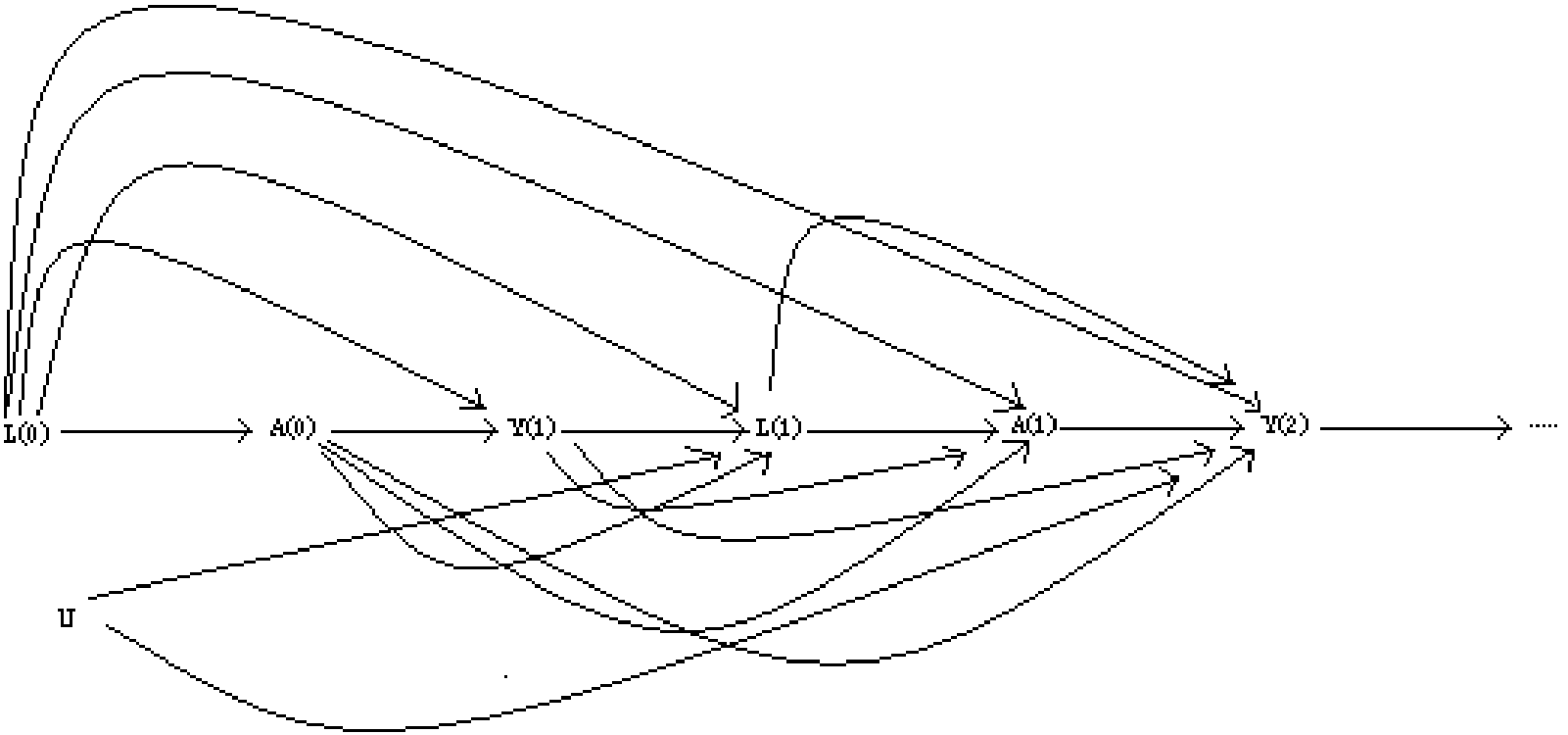
- The Multicenter Aids Cohort Study (MACS) followed HIV+ men in 4 US cities,
- We restict our attention to data from1996 to 2002,that is the HAART era
- Data includes the following covariates measured at baseline and at every bi-annual visit:HAART use, CD4 cell count, plasma HIV-1 viral load,use of mono-therapy or combination antiretroviral therapy.
- Goal: Estimate the causal effect of HAART on CD4 count evolution

Causal Inference for Longitudinal Data

- Write $A(j)$ as HAART use at time j, $L(j)$ vector of all confounders measured at time j, $Y(j)$ CD4 cell count measure at time j.

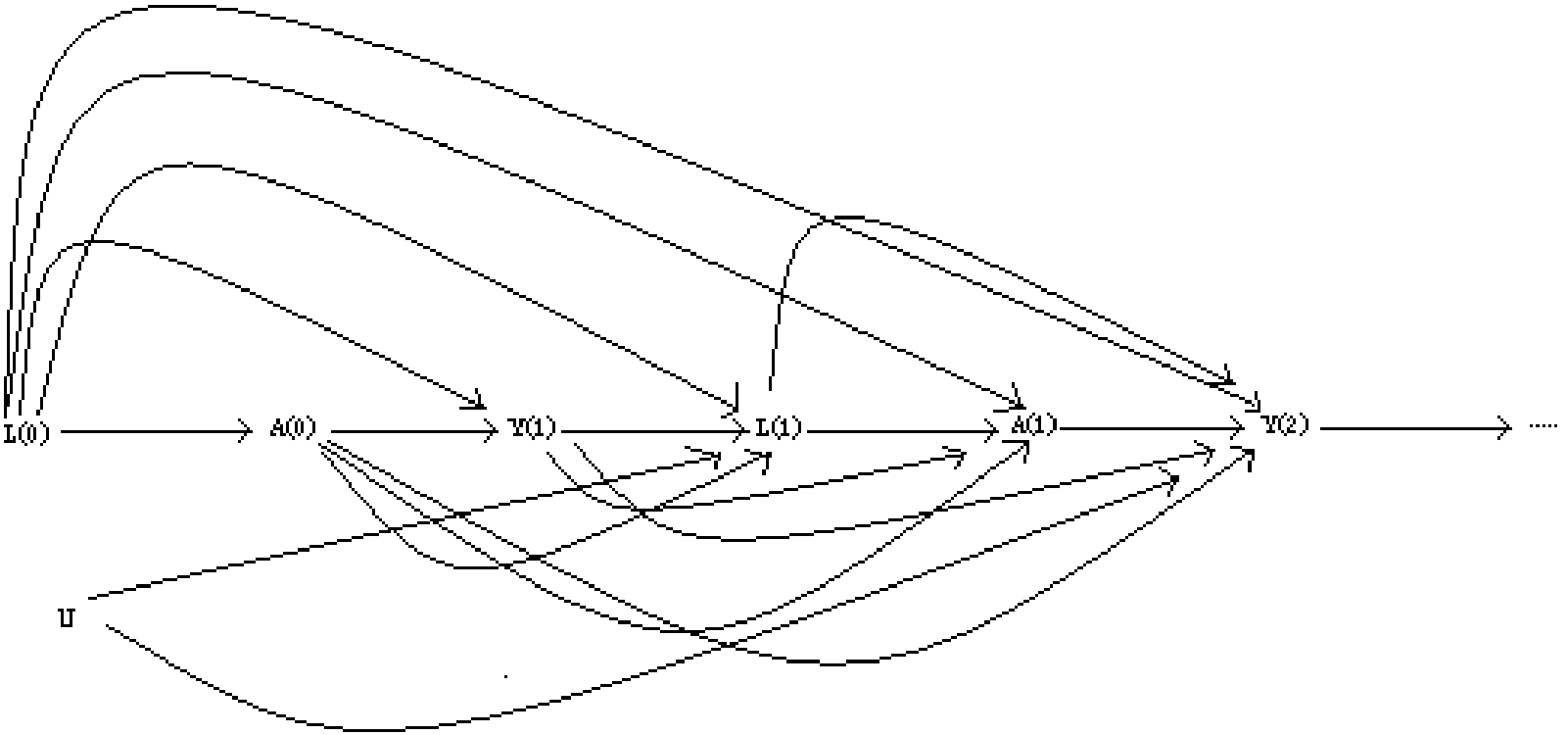
Causal Inference for Longitudinal Data

- Write $A(j)$ as HAART use at time j, $L(j)$ vector of all confounders measured at time j, $Y(j)$ CD4 cell count measure at time j.
- The causal DAG is of the form:



Causal Inference for Longitudinal Data

- Write $A(j)$ as HAART use at time j, $L(j)$ vector of all confounders measured at time j, $Y(j)$ CD4 cell count measure at time j.
- The causal DAG is of the form:



- U is an unmeasured common cause of some components of $L(1)$ and $Y(2)$.

Causal Inference for Longitudinal Data

- A standard way to check whether say $A(0)$ and $A(1)$ has a direct effect on $Y(2)$ controlling for confounding is to use a regression model of the form:

$$\begin{aligned} & E\left(Y(2)\left|A(0), A(1), L(0), L(1), Y(1)\right.\right) \\ = \quad & \beta_0 + \left(L'(0), L'(1), Y(1)\right)\beta_1 + \beta_2 A(0) + \beta_3 A(1) \end{aligned} \tag{2}$$

Causal Inference for Longitudinal Data

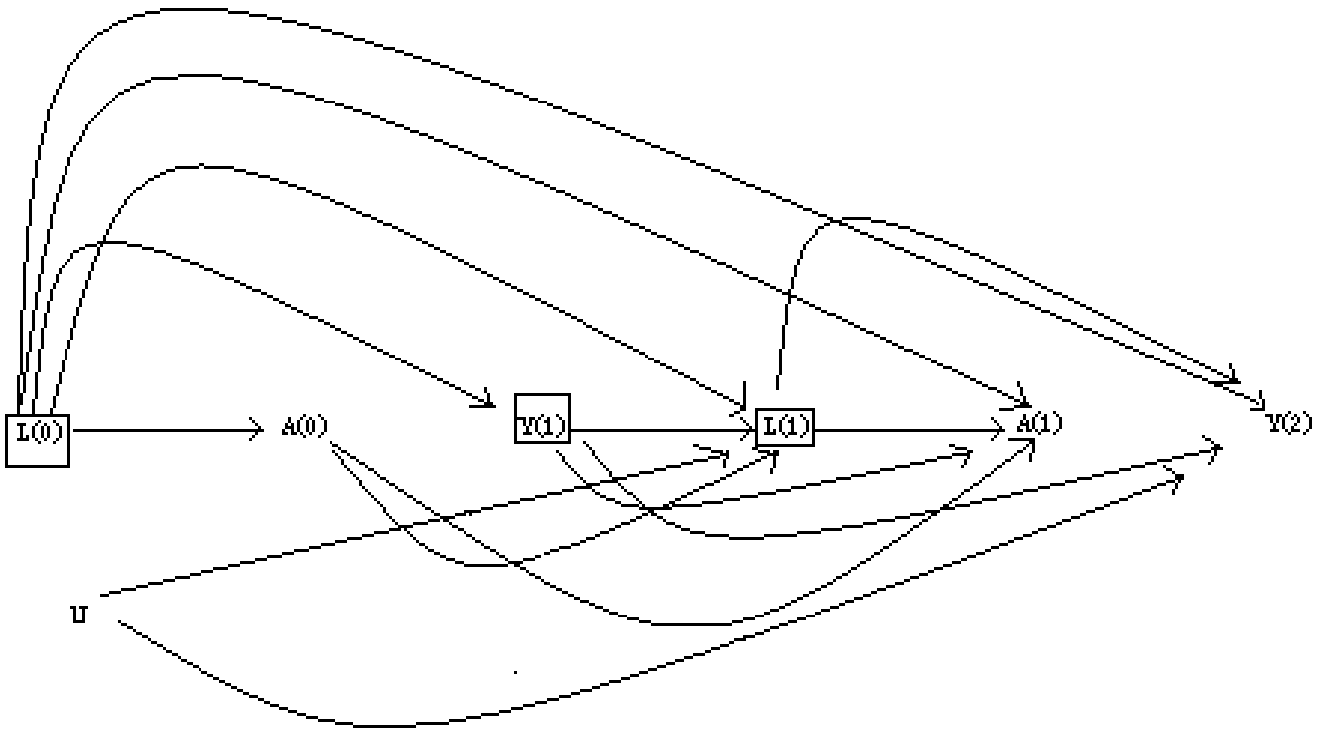
- A standard way to check whether say $A(0)$ and $A(1)$ has a direct effect on $Y(2)$ controlling for confounding is to use a regression model of the form:

$$\begin{aligned} & E\left(Y(2)\left|A(0), A(1), L(0), L(1), Y(1)\right.\right) \\ = \quad & \beta_0 + \left(L'(0), L'(1), Y(1)\right)\beta_1 + \beta_2 A(0) + \beta_3 A(1) \end{aligned} \tag{2}$$

- Here we assume that model misspecification is absent.

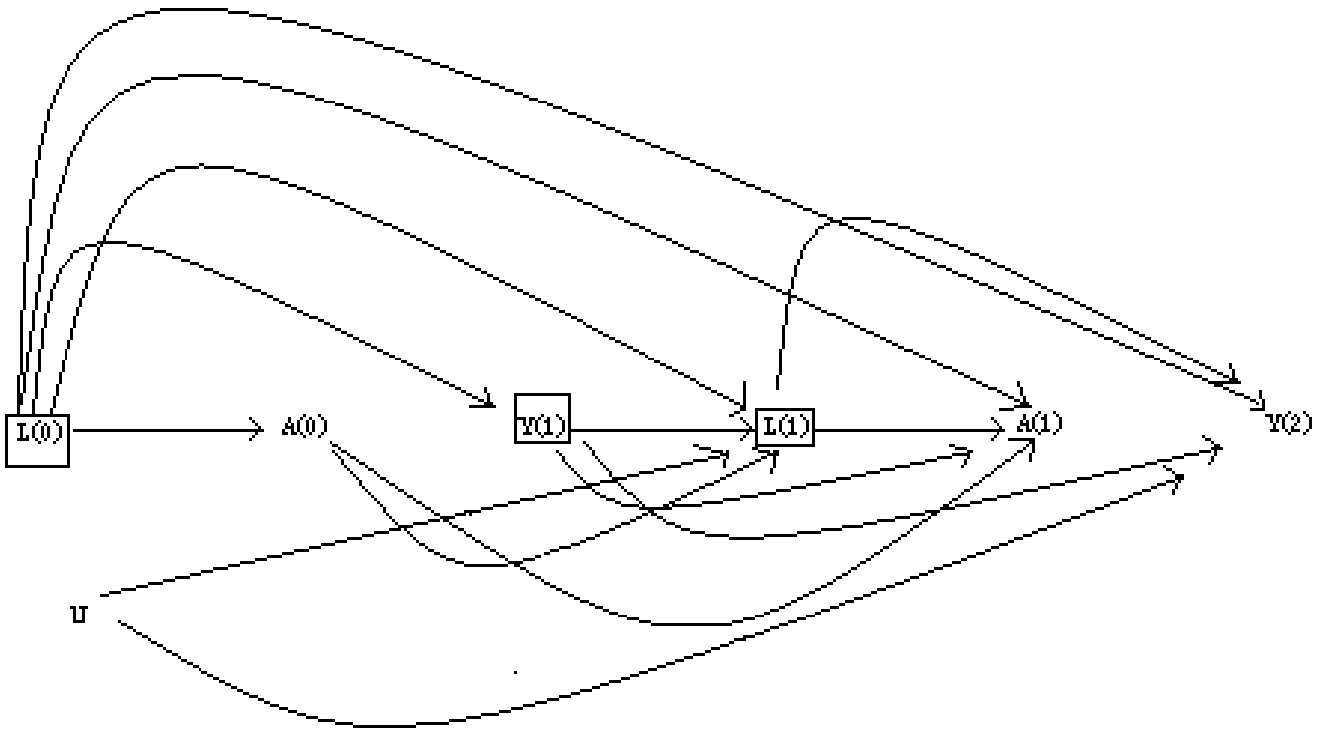
Causal Inference for Longitudinal Data

- To reveal the pitfall of such a standard analysis, lets assume we are under the joint null that A has no direct effect on Y holds, Then, model (2) is equivalent to the following DAG



Causal Inference for Longitudinal Data

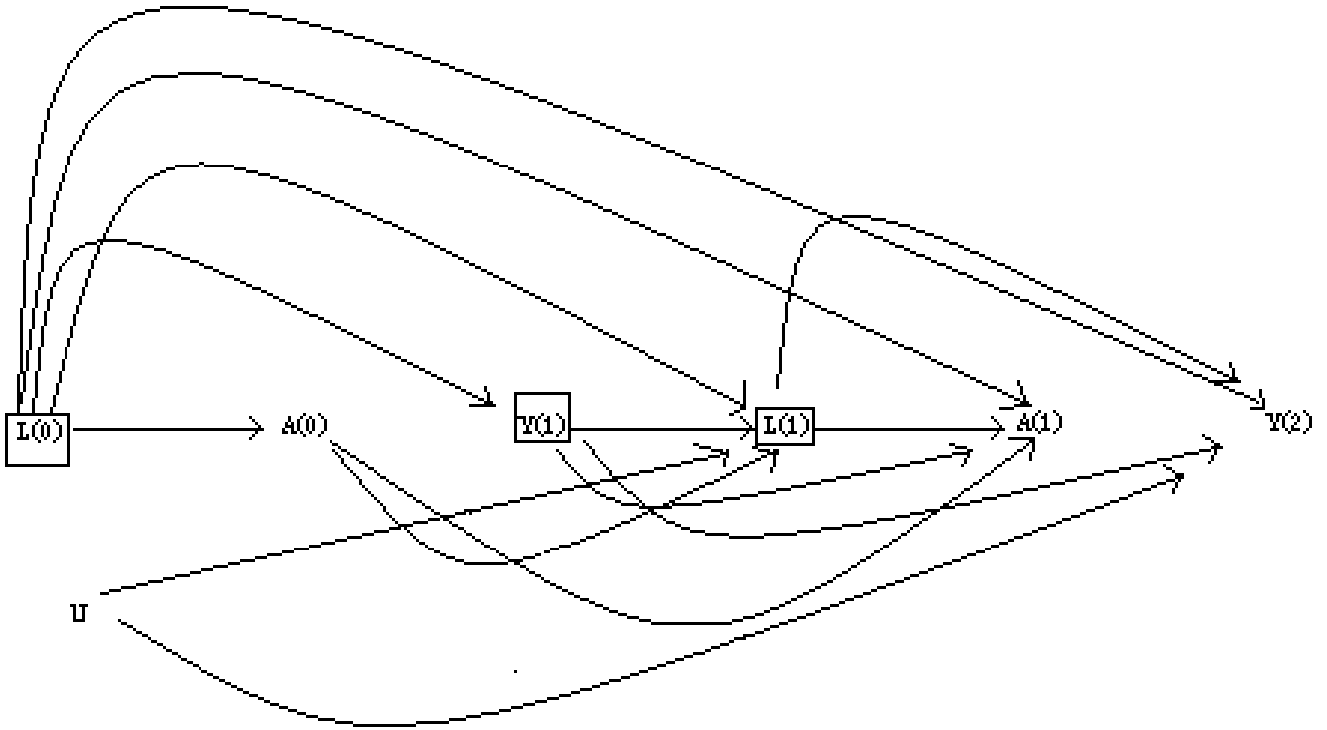
- To reveal the pitfall of such a standard analysis, lets assume we are under the joint null that A has no direct effect on Y holds, Then, model (2) is equivalent to the following DAG



- Where as consistent with the causal null, we have dropped the directed arrows:$A(0) \rightarrow Y(1), A(0) \rightarrow Y(2), A(1) \rightarrow Y(2)$ and so on ...

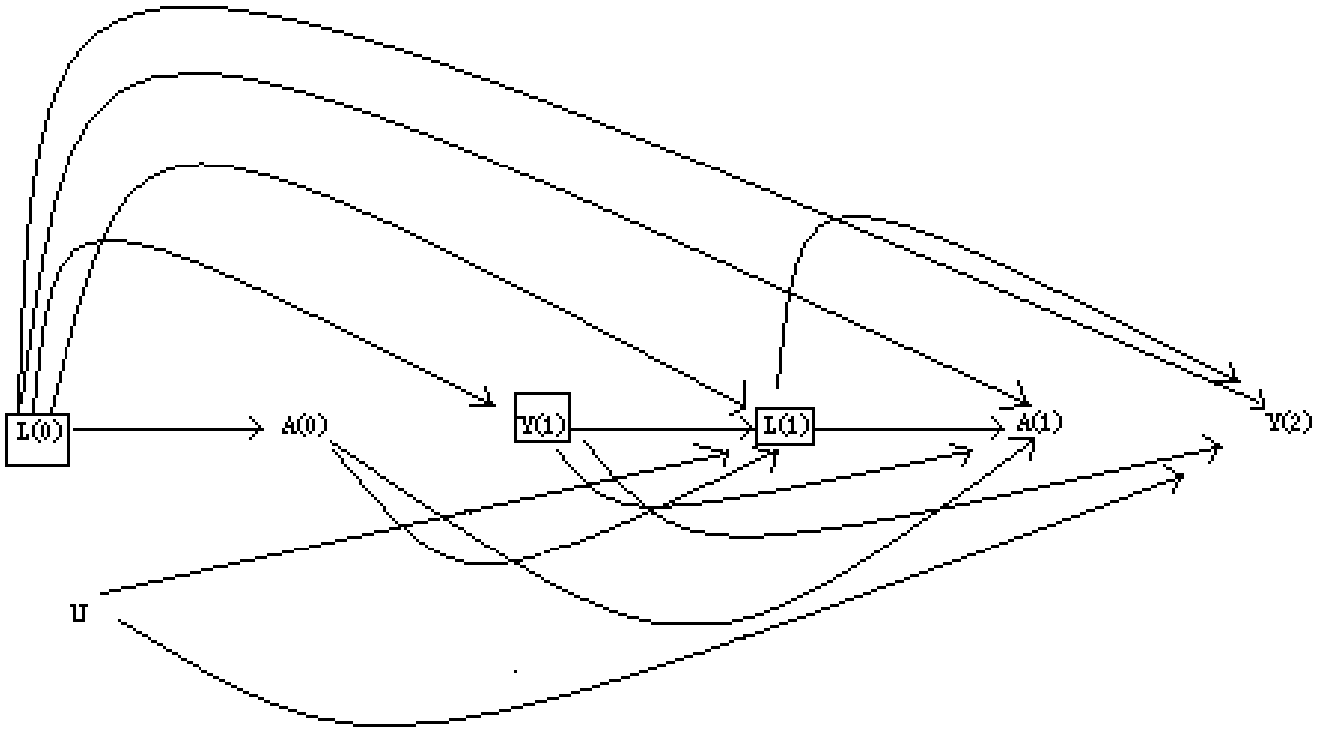
Causal Inference for Longitudinal Data

- But by our theory of DAGs, we see that we are conditioning on $L(1)$ which is a collider along the path A(0)-L(1)-U-Y(2),



Causal Inference for Longitudinal Data

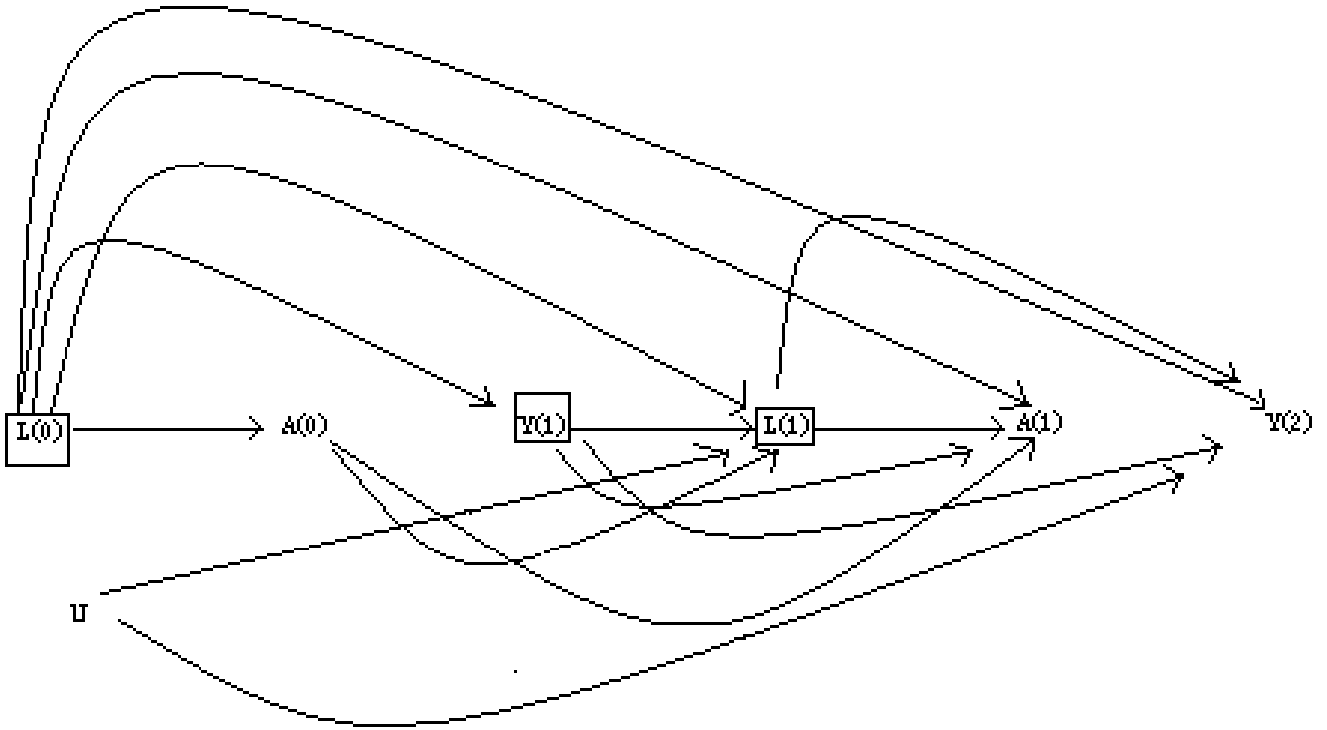
- But by our theory of DAGs, we see that we are conditioning on $L(1)$ which is a collider along the path A(0)-L(1)-U-Y(2),



- Thus A(0) is correlated with Y(2) given L(0), Y(1),L(1) and A(1)$\Rightarrow \beta_2 \neq 0$ in model (2), which contradicts the null of no causal effect.

Causal Inference for Longitudinal Data

- But by our theory of DAGs, we see that we are conditioning on $L(1)$ which is a collider along the path A(0)-L(1)-U-Y(2),



- Thus A(0) is correlated with Y(2) given L(0), Y(1),L(1) and A(1)$\Rightarrow \beta_2 \neq 0$ in model (2), which contradicts the null of no causal effect.
- We conclude that β_2 cannot logically have a causal interpretation,

Causal Inference for Longitudinal Data

- In our example, conditioning on $L(1)$ caused the problem, you might consider a model (3)

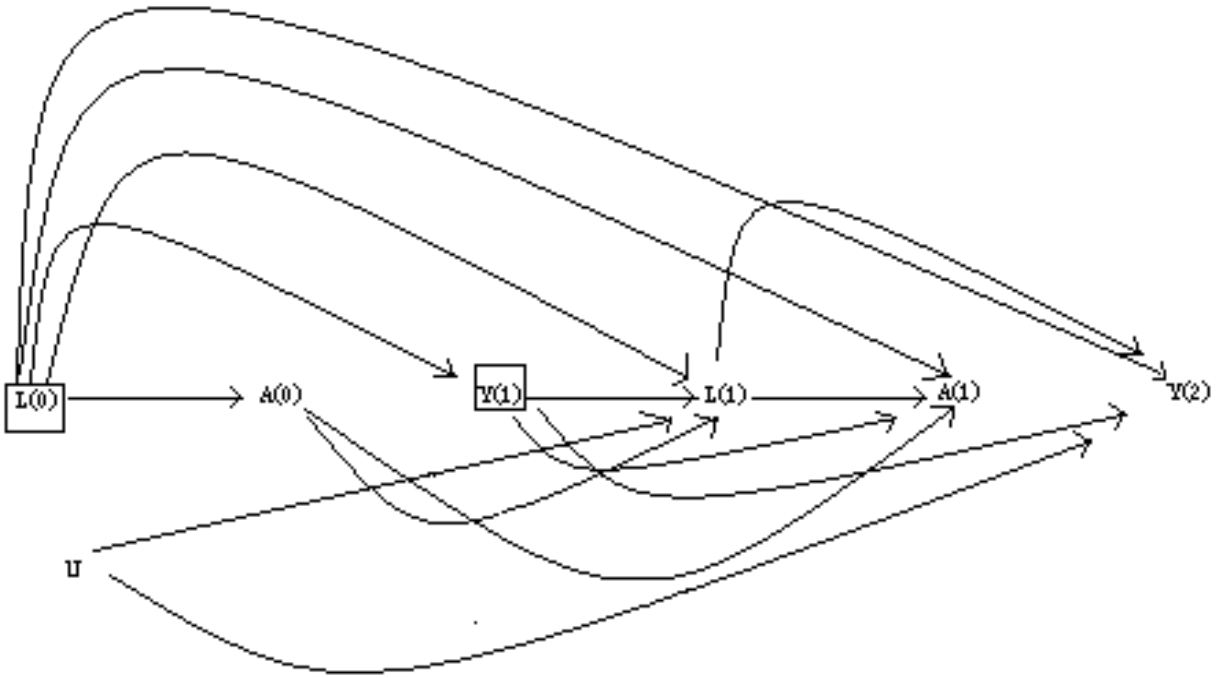
$$\begin{aligned} & E\left(Y\left(2\right)|A\left(0\right),A\left(1\right),L\left(0\right),Y\left(1\right)\right) \\ = \quad & \beta_0+\left(L'\left(0\right),Y\left(1\right)\right)\beta_1+\beta_2 A\left(0\right)+\beta_3 A\left(1\right) \end{aligned} \tag{3}$$

Causal Inference for Longitudinal Data

- In our example, conditioning on $L(1)$ caused the problem, you might consider a model (3)

$$\begin{aligned} & E\left(Y\left(2\right)|A\left(0\right),A\left(1\right),L\left(0\right),Y\left(1\right)\right) \\ = \quad & \beta_0+\left(L'\left(0\right),Y\left(1\right)\right)\beta_1+\beta_2 A\left(0\right)+\beta_3 A\left(1\right) \end{aligned} \tag{3}$$

- But this corresponds to the DAG



Causal Inference for Longitudinal Data

- Again by our theory of DAGs, we have an open backdoor path A(1)-L(1)-Y(2)

Causal Inference for Longitudinal Data

- Again by our theory of DAGs, we have an open backdoor path A(1)-L(1)-Y(2)
- So that $\beta_3 \neq 0$, and again we would wrongly conclude that A causes Y.

Causal Inference for Longitudinal Data

- Again by our theory of DAGs, we have an open backdoor path A(1)-L(1)-Y(2)
- So that $\beta_3 \neq 0$, and again we would wrongly conclude that A causes Y.
- Thus now, β_3 cannot logically have a causal interpretation.

Causal Inference for Longitudinal Data

- Again by our theory of DAGs, we have an open backdoor path A(1)-L(1)-Y(2)
- So that $\beta_3 \neq 0$, and again we would wrongly conclude that A causes Y.
- Thus now, β_3 cannot logically have a causal interpretation.
- The problem is that no parameter in model (2) clearly represents the causal effect of A on Y.

Causal Inference for Longitudinal Data

- Again by our theory of DAGs, we have an open backdoor path A(1)-L(1)-Y(2)
- So that $\beta_3 \neq 0$, and again we would wrongly conclude that A causes Y.
- Thus now, β_3 cannot logically have a causal interpretation.
- The problem is that no parameter in model (2) clearly represents the causal effect of A on Y.
- A new approach is needed.

Causal Inference for Longitudinal Data

- First we must define our causal model in this setting.

Causal Inference for Longitudinal Data

- First we must define our causal model in this setting.
- Consider Y(2) again, and define $Y_{a_0,a_1}(2)$ the participant's counterfactual CD4 count at the second occasion had he possibly contrary to fact received treatment history $A_0 = a_0, A_1 = a_1$.

Causal Inference for Longitudinal Data

- First we must define our causal model in this setting.
- Consider Y(2) again, and define $Y_{a_0,a_1}(2)$ the participant's counterfactual CD4 count at the second occasion had he possibly contrary to fact received treatment history $A_0 = a_0, A_1 = a_1$.
- In general, we will use $E\left(Y_{a_0,a_1}(2)\right) = g\left(a_0, a_1\right)$ to capture the joint effect of $A(1)$ and $A(2)$.

Causal Inference for Longitudinal Data

- First we must define our causal model in this setting.
- Consider Y(2) again, and define $Y_{a_0,a_1}(2)$ the participant's counterfactual CD4 count at the second occasion had he possibly contrary to fact received treatment history $A_0 = a_0, A_1 = a_1$.
- In general, we will use $E\left(Y_{a_0,a_1}(2)\right) = g\left(a_0, a_1\right)$ to capture the joint effect of $A(1)$ and $A(2)$.
- e.g. $E\left(Y_{1,1}(2)\right) - E\left(Y_{0,0}(2)\right)$ is the causal difference in the mean CD4 count when on HAART at occasions 1 & 2 versus never receiving HAART.

Causal Inference for Longitudinal Data

- $g\left(a_0, a_1\right)$ is a counterfactual mean, under what assumptions can it be identified from the observed data?

Causal Inference for Longitudinal Data

- $g\left(a_0, a_1\right)$ is a counterfactual mean, under what assumptions can it be identified from the observed data?
- We give a generalization of the no unmeasured confounder assumption, also known as *sequential randomization* with respect to $Y_{a_0,a_1}(2)$.

$$\left\{Y_{a_0,a_1}(2) : a_0, a_1\right\} \amalg A(0) \left| L(0) \right. \text{ and}$$
$$\left\{Y_{a_0,a_1}(2) : a_0, a_1\right\} \amalg A(1) \left| A(0), L(0), Y(1), L(1) \right.$$

Causal Inference for Longitudinal Data

- $g\left(a_0, a_1\right)$ is a counterfactual mean, under what assumptions can it be identified from the observed data?
- We give a generalization of the no unmeasured confounder assumption, also known as *sequential randomization* with respect to $Y_{a_0,a_1}(2)$.

$$\left\{Y_{a_0,a_1}(2) : a_0, a_1\right\} \amalg A(0) \left| L(0) \right. \text{ and}$$
$$\left\{Y_{a_0,a_1}(2) : a_0, a_1\right\} \amalg A(1) \left| A(0), L(0), Y(1), L(1) \right.$$

- This assumption essentially says that at $A(0)$ is randomized within levels of $L(0)$ and $A(1)$ is randomized within the joint levels of $A(0), L(0), Y(1), L(1)$.

Causal Inference for Longitudinal Data

- We also make the consistency assumption: $Y(2) = Y_{A_0,A_1}(2)$ w.p.1

Causal Inference for Longitudinal Data

- We also make the consistency assumption: $Y(2) = Y_{A_0,A_1}(2)$ w.p.1
- The g-formula of Robins states that under these assumptions:

$$\begin{aligned} & E\left(Y_{a_0,a_1}(2)\right) \\ = \; & g\left(a_0, a_1\right) \\ = \; & \iiint E\left(Y|A(0)=a_0, A(1)=a_1, Y(1)=y_1, L_1=l_1, L_0=l_0\right) \\ & \times f_{L_1,Y(1)|A(0),L_0}\left(l_1, Y(1)=y_1|A(0)=a_0, L_0=l_0\right) \\ & \times f_{L_0}\left(l_0\right) dl_0 dY(1) dl_1 \end{aligned}$$

- Proof:

$$
\begin{aligned}
& E\left(Y_{a_0,a_1}(2)\right) \\
= & \int E\left(Y_{a_0,a_1}(2)|L_0=l_0\right) f_{L_0}\left(l_0\right) dl_0 \\
= & \int E\left(Y_{a_0,a_1}(2)|L_0=l_0, A_0=a_0\right) f_{L_0}\left(l_0\right) dl_0 \\
= & \iint E\left(Y_{a_0,a_1}(2)|L_0=l_0, A_0=a_0, Y\left(1\right)=y_1, L_1=l_1\right) \\
& \times f_{L_1,Y(1)|A(0),L_0}\left(l_1,y_1|A\left(0\right)=a_0, L_0=l_0\right) f_{L_0}\left(l_0\right) dl_0 dY(1) dl_1 \\
= & \iint E\left(Y(2)|A\left(0\right)=a_0, A\left(1\right)=a_1, L_1=l_1, Y\left(1\right)=y_1, L_0=l_0\right) \\
& \times f_{L_1,Y(1)|A(0),L_0}\left(l_1,y_1|A\left(0\right)=a_0, L_0=l_0\right) f_{L_0}\left(l_0\right) dl_0 dY(1) dl_1
\end{aligned}
$$

Causal Inference for Longitudinal Data

- The $G-$ *formula* can be generalized to an arbitrary number of repeated measures.

Causal Inference for Longitudinal Data

- The $G-$ *formula* can be generalized to an arbitrary number of repeated measures.
- As before, G-computation of $E\left(Y_{a_0,a_1}(2)\right)$, requires estimates of

Causal Inference for Longitudinal Data

- The $G-$ *formula* can be generalized to an arbitrary number of repeated measures.
- As before, G-computation of $E\left(Y_{a_0,a_1}(2)\right)$, requires estimates of
  - $E\left(Y(2)|A(0)=a_0, A(1)=a_1, L_1=l_1, Y(1)=y_1, L_0=l_0\right)$,

Causal Inference for Longitudinal Data

- The $G-$ *formula* can be generalized to an arbitrary number of repeated measures.
- As before, G-computation of $E\left(Y_{a_0,a_1}(2)\right)$, requires estimates of
  - $E\left(Y(2)|A(0)=a_0, A(1)=a_1, L_1=l_1, Y(1)=y_1, L_0=l_0\right)$,
  - $f_{L_1,Y(1)|A(0),L_0}\left(l_1, y_1|A(0)=a_0, L_0=l_0\right)$

Causal Inference for Longitudinal Data

- The $G-$ *formula* can be generalized to an arbitrary number of repeated measures.
- As before, G-computation of $E\left(Y_{a_0,a_1}(2)\right)$, requires estimates of
  - $E\left(Y(2)|A(0)=a_0, A(1)=a_1, L_1=l_1, Y(1)=y_1, L_0=l_0\right)$,
  - $f_{L_1,Y(1)|A(0),L_0}\left(l_1, y_1|A(0)=a_0, L_0=l_0\right)$
  - $f_{L_0}\left(l_0\right)$

Causal Inference for Longitudinal Data

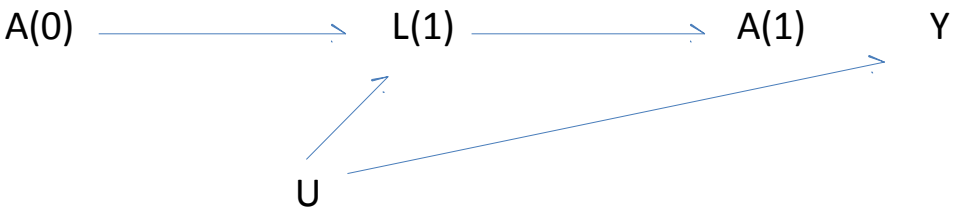
- The $G-$ *formula* can be generalized to an arbitrary number of repeated measures.
- As before, G-computation of $E\left(Y_{a_0,a_1}(2)\right)$, requires estimates of
  - $E\left(Y(2)|A(0)=a_0, A(1)=a_1, L_1=l_1, Y(1)=y_1, L_0=l_0\right)$,
  - $f_{L_1,Y(1)|A(0),L_0}\left(l_1, y_1|A(0)=a_0, L_0=l_0\right)$
  - $f_{L_0}\left(l_0\right)$
- It will generally be difficult to specify models for these quantities,such that the g-formula is not mis-specified.

Causal Inference for Longitudinal Data

- The $G-$ *formula* can be generalized to an arbitrary number of repeated measures.
- As before, G-computation of $E\left(Y_{a_0,a_1}(2)\right)$, requires estimates of
  - $E\left(Y(2)|A(0)=a_0, A(1)=a_1, L_1=l_1, Y(1)=y_1, L_0=l_0\right)$,
  - $f_{L_1,Y(1)|A(0),L_0}\left(l_1, y_1|A(0)=a_0, L_0=l_0\right)$
  - $f_{L_0}\left(l_0\right)$
- It will generally be difficult to specify models for these quantities,such that the g-formula is not mis-specified.
- The main issue is that under standard parametrization, there is no parameter to encode the null hypothesis of no joint effect of (a_0, a_1).

Null Paradox

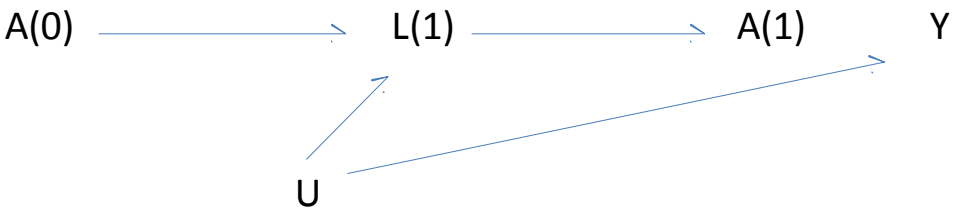
- For instance, consider the causal graph



under the Null hypothesis of no joint effect of $(A(0), A(1))$ on Y

Null Paradox

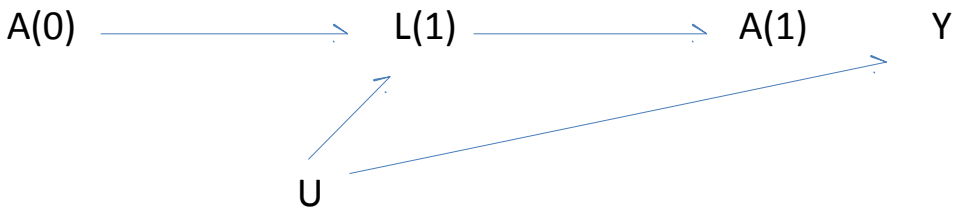
- For instance, consider the causal graph



  under the Null hypothesis of no joint effect of $(A(0), A(1))$ on Y
- Suppose that $L(1)$ is binary and Y is continuous, so that the g-formula in this graph gives $E\left(Y_{a_0,a_1}\right) = \sum_{l_1=0}^{1} E\left(Y|A(0) = a_0, L(1) = l_1\right) \times f_{L(1)|A(0)}\ \left(l_1|A\left(0\right) = a_0\right) dl_1$

Null Paradox

- For instance, consider the causal graph



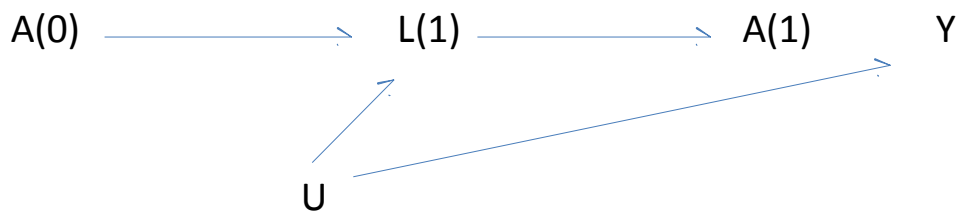
  under the Null hypothesis of no joint effect of $(A(0), A(1))$ on Y
- Suppose that $L(1)$ is binary and Y is continuous, so that the g-formula in this graph gives $E\left(Y_{a_0,a_1}\right) = \sum_{l_1=0}^{1} E\left(Y|A(0)=a_0, L(1)=l_1\right) \times f_{L(1)|A(0)}\ \left(l_1|A\left(0\right)=a_0\right) dl_1$
- A standard modeling approach would fit a linear regression

$$E\left(Y|A_0 = a_0, A(1) = a_1, L(1) = l_1; \gamma\right) = \left(1, a_0, a_1, l_1\right)\gamma$$

  and a logistic regression

$$\text{logit}\,\Pr(L(1) = 1|A(0) = 0; \alpha) = (1, a_0)\alpha$$

Null Paradox



The estimated causal effect

$$\begin{aligned}
& \widehat{E}\left(Y_{a_0,a_1}\right) \\
= & \sum_{l_1=0}^{1}\left[E\left(Y|A(0)=a_0, A(1)=a_1, L(1)=l_1;\widehat{\gamma}\right)\right. \\
& \left.\times f_{L(1)|A(0)}\ \left(l_1|A\left(0\right)=a_0;\widehat{\alpha}\right)\right] \\
= & \left(1, a_0,\ \ a_1, \frac{\exp\left((1,a_0)\widehat{\alpha}\right)}{1+\exp\left((1,a_0)\widehat{\alpha}\right)}\right)\widehat{\gamma}
\end{aligned}$$

therefore $\widehat{E}\left(Y_{a_0,a_1}\right) \xrightarrow{P} E\left(Y_{a_0,a_1}\right)$ does not depend on $\left(a_0, a_1\right)$ if either $\gamma_1 = \gamma_2 = \gamma_3 = 0$ or $\gamma_1 = \gamma_2 = \alpha_2 = 0$

Null Paradox

- Therefore $\widehat{E}\left(Y_{a_0,a_1}\right) \xrightarrow{P} E\left(Y_{a_0,a_1}\right)$ does not depend on $\left(a_0, a_1\right)$ if either $\gamma_1 = \gamma_2 = \gamma_3 = 0$ or $\gamma_1 = \gamma_2 = \alpha_2 = 0$

Null Paradox

- Therefore $\widehat{E}\left(Y_{a_0,a_1}\right) \xrightarrow{P} E\left(Y_{a_0,a_1}\right)$ does not depend on $\left(a_0, a_1\right)$ if either $\gamma_1 = \gamma_2 = \gamma_3 = 0$ or $\gamma_1 = \gamma_2 = \alpha_2 = 0$
- However $\gamma_3 \neq 0$ because $L(1)$ is correlated with Y

Null Paradox

- Therefore $\widehat{E}\left(Y_{a_0,a_1}\right) \xrightarrow{P} E\left(Y_{a_0,a_1}\right)$ does not depend on (a_0, a_1) if either $\gamma_1 = \gamma_2 = \gamma_3 = 0$ or $\gamma_1 = \gamma_2 = \alpha_2 = 0$
- However $\gamma_3 \neq 0$ because $L(1)$ is correlated with Y
- $\alpha_2 \neq 0$ because $A(0)$ predicts $L(1)$

Null Paradox

- Therefore $\widehat{E}\left(Y_{a_0,a_1}\right) \xrightarrow{P} E\left(Y_{a_0,a_1}\right)$ does not depend on (a_0, a_1) if either $\gamma_1 = \gamma_2 = \gamma_3 = 0$ or $\gamma_1 = \gamma_2 = \alpha_2 = 0$
- However $\gamma_3 \neq 0$ because $L(1)$ is correlated with Y
- $\alpha_2 \neq 0$ because $A(0)$ predicts $L(1)$
- If causal null hypothesis is true, then we know model is misspecified. . . before we collect any data!

Null Paradox

- Therefore $\widehat{E}\left(Y_{a_0,a_1}\right) \xrightarrow{P} E\left(Y_{a_0,a_1}\right)$ does not depend on (a_0, a_1) if either $\gamma_1 = \gamma_2 = \gamma_3 = 0$ or $\gamma_1 = \gamma_2 = \alpha_2 = 0$
- However $\gamma_3 \neq 0$ because $L(1)$ is correlated with Y
- $\alpha_2 \neq 0$ because $A(0)$ predicts $L(1)$
- If causal null hypothesis is true, then we know model is misspecified. . . before we collect any data!
- We need other kinds of models

Causal Inference for Longitudinal Data

- An alternative is to directly specify a model for the marginal mean

$$E\left(Y_{a_0,a_1}(2);\psi\right)$$

with finite dimensional parameter ψ.

Causal Inference for Longitudinal Data

- An alternative is to directly specify a model for the marginal mean

$$E\left(Y_{a_0,a_1}(2);\psi\right)$$

  with finite dimensional parameter ψ.
- For example, we could specify:

Causal Inference for Longitudinal Data

- An alternative is to directly specify a model for the marginal mean

$$E\left(Y_{a_0,a_1}(2);\psi\right)$$

  with finite dimensional parameter ψ.
- For example, we could specify:
  - $E\left(Y_{a_0,a_1}(2);\psi\right) = \psi_0 + \psi_1 a_0 + \psi_2 a_1$ or

Causal Inference for Longitudinal Data

- An alternative is to directly specify a model for the marginal mean

$$E\left(Y_{a_0,a_1}(2);\psi\right)$$

  with finite dimensional parameter ψ.
- For example, we could specify:
  - $E\left(Y_{a_0,a_1}(2);\psi\right) = \psi_0 + \psi_1 a_0 + \psi_2 a_1$ or
  - $E\left(Y_{a_0,a_1}(2);\psi\right) = \psi_0 + \psi_1\left(a_0 + a_1\right)$ or

Causal Inference for Longitudinal Data

- An alternative is to directly specify a model for the marginal mean

$$E\left(Y_{a_0,a_1}(2);\psi\right)$$

  with finite dimensional parameter ψ.
- For example, we could specify:
  - $E\left(Y_{a_0,a_1}(2);\psi\right) = \psi_0 + \psi_1 a_0 + \psi_2 a_1$ or
  - $E\left(Y_{a_0,a_1}(2);\psi\right) = \psi_0 + \psi_1 \left(a_0 + a_1\right)$ or
  - $E\left(Y_{a_0,a_1}(2);\psi\right) = \psi_0 + \psi_1 a_1$

Causal Inference for Longitudinal Data

- An alternative is to directly specify a model for the marginal mean

$$E\left(Y_{a_0,a_1}(2);\psi\right)$$

  with finite dimensional parameter ψ.
- For example, we could specify:
  - $E\left(Y_{a_0,a_1}(2);\psi\right) = \psi_0 + \psi_1 a_0 + \psi_2 a_1$ or
  - $E\left(Y_{a_0,a_1}(2);\psi\right) = \psi_0 + \psi_1 \left(a_0 + a_1\right)$ or
  - $E\left(Y_{a_0,a_1}(2);\psi\right) = \psi_0 + \psi_1 a_1$
- Note here that ψ has a causal interpretation, it is the parameter of a *Marginal Structural Mean Model (*MSMM)

Causal Inference for Longitudinal Data

- An alternative is to directly specify a model for the marginal mean

$$E\left(Y_{a_0,a_1}(2);\psi\right)$$

  with finite dimensional parameter ψ.
- For example, we could specify:
  - $E\left(Y_{a_0,a_1}(2);\psi\right) = \psi_0 + \psi_1 a_0 + \psi_2 a_1$ or
  - $E\left(Y_{a_0,a_1}(2);\psi\right) = \psi_0 + \psi_1 \left(a_0 + a_1\right)$ or
  - $E\left(Y_{a_0,a_1}(2);\psi\right) = \psi_0 + \psi_1 a_1$
- Note here that ψ has a causal interpretation, it is the parameter of a *Marginal Structural Mean Model (*MSMM)
- So that $E\left(Y_{a_0,a_1}(2);\psi\right) = E\left(Y_{0,0}(2);\psi\right) \Leftrightarrow \psi_1 = \psi_2 = 0$

Causal Inference for Longitudinal Data

- The parameters of a MSMM, say $E\left(Y_{a_0,a_1}(2);\psi\right) = \psi_0 + \psi_1 a_0 + \psi_2 a_1$ cannot be estimated by OLS

Causal Inference for Longitudinal Data

- The parameters of a MSMM, say $E\left(Y_{a_0,a_1}(2);\psi\right) = \psi_0 + \psi_1 a_0 + \psi_2 a_1$ cannot be estimated by OLS
- We need a generalization of iptw:

Causal Inference for Longitudinal Data

- The parameters of a MSMM, say $E\left(Y_{a_0,a_1}(2);\psi\right) = \psi_0 + \psi_1 a_0 + \psi_2 a_1$ cannot be estimated by OLS
- We need a generalization of iptw:
  - Specify models $\pi\left(L_0;\alpha_0\right)$ and $\pi\left(L_1, A_0, L_0, Y\left(1\right);\alpha_1\right)$for $f\left(A_0|L_0\right)$ and $f\left(A_1|L_1, A_0, L_0, Y\left(1\right)\right)$; say logistic regressions and obtain the MLEs $\widehat{\alpha}_0$ and $\widehat{\alpha}_1$

Causal Inference for Longitudinal Data

- The parameters of a MSMM, say $E\left(Y_{a_0,a_1}(2);\psi\right) = \psi_0 + \psi_1 a_0 + \psi_2 a_1$ cannot be estimated by OLS
- We need a generalization of iptw:
  - Specify models $\pi\left(L_0;\alpha_0\right)$ and $\pi\left(L_1, A_0, L_0, Y\left(1\right);\alpha_1\right)$for $f\left(A_0|L_0\right)$ and $f\left(A_1|L_1, A_0, L_0, Y\left(1\right)\right)$; say logistic regressions and obtain the MLEs $\widehat{\alpha}_0$ and $\widehat{\alpha}_1$
  - For each person in the study, compute the weight $W = f\left(A_0|L_0;\widehat{\alpha}_0\right) f\left(A_1|L_1, A_0, L_0, Y\left(1\right);\widehat{\alpha}_1\right)$ which corresponds to the estimated probability of receiving the treatment you did indeed receive.

Causal Inference for Longitudinal Data

- The parameters of a MSMM, say $E\left(Y_{a_0,a_1}(2);\psi\right)=\psi_0+\psi_1 a_0+\psi_2 a_1$ cannot be estimated by OLS
- We need a generalization of iptw:
  - Specify models $\pi\left(L_0;\alpha_0\right)$ and $\pi\left(L_1,A_0,L_0,Y\left(1\right);\alpha_1\right)$for $f\left(A_0|L_0\right)$ and $f\left(A_1|L_1,A_0,L_0,Y\left(1\right)\right)$; say logistic regressions and obtain the MLEs $\widehat{\alpha}_0$ and $\widehat{\alpha}_1$
  - For each person in the study, compute the weight $W=f\left(A_0|L_0;\widehat{\alpha}_0\right)f\left(A_1|L_1,A_0,L_0,Y\left(1\right);\widehat{\alpha}_1\right)$ which corresponds to the estimated probability of receiving the treatment you did indeed receive.
  - Regress $Y(2)$ on A_0 and A_1 using weighted least-squares (WLS) with weights W^{-1}, $\widehat{\psi}$ is the iptw estimate of ψ.

Causal Inference for Longitudinal Data

- As long as the weights are consistently estimated, $\widehat{\psi}$ is consistent for ψ.

Causal Inference for Longitudinal Data

- As long as the weights are consistently estimated, $\widehat{\psi}$ is consistent for ψ.
- Easy to extend this approach to most standard statistical models, including repeated measures and survival outcomes.

Causal Inference for Longitudinal Data

- As long as the weights are consistently estimated, $\widehat{\psi}$ is consistent for ψ.
- Easy to extend this approach to most standard statistical models, including repeated measures and survival outcomes.
- To account for the preliminary estimation of the weights, we require the use of the so-called *robust variance*

Causal Inference for Longitudinal Data

- As long as the weights are consistently estimated, $\widehat{\psi}$ is consistent for ψ.
- Easy to extend this approach to most standard statistical models, including repeated measures and survival outcomes.
- To account for the preliminary estimation of the weights, we require the use of the so-called *robust variance*
- It is advisable to stabilize the weights by using $SW = \frac{f(A_0|L_0;\widehat{\alpha}_0)f(A_1|L_1,A_0,L_0,Y(1);\widehat{\alpha}_1)}{f(A_0|\widehat{\rho}_0)f(A_1|A_0;\widehat{\rho}_1)}$ instead of W, where $f\left(A_0|\widehat{\rho}_0\right)$ and $f\left(A_1|A_0;\widehat{\rho}_1\right)$ are estimated using finite dimensional logistic regression models.

Causal Inference for Longitudinal Data

- As long as the weights are consistently estimated, $\widehat{\psi}$ is consistent for ψ.
- Easy to extend this approach to most standard statistical models, including repeated measures and survival outcomes.
- To account for the preliminary estimation of the weights, we require the use of the so-called *robust variance*
- It is advisable to stabilize the weights by using $SW = \frac{f(A_0|L_0;\widehat{\alpha}_0)f(A_1|L_1,A_0,L_0,Y(1);\widehat{\alpha}_1)}{f(A_0|\widehat{\rho}_0)f(A_1|A_0;\widehat{\rho}_1)}$ instead of W, where $f\left(A_0|\widehat{\rho}_0\right)$ and $f\left(A_1|A_0;\widehat{\rho}_1\right)$ are estimated using finite dimensional logistic regression models.
- dr estimation methods also available to partially protect against misspecified weights