Statistical Learning
An Introduction to Causal Inference

Introduction

There are generally two notions of causation:

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

Our causal paradigm consists of:

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

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:
  Y^0: headache outcome after not taking aspirin
  Y^1: headache outcome after taking aspirin

Counterfactuals

- Y^a is the outcome that you would observe if possibly countering to fact you followed treatment $a \in \{0, 1\}$.
- A mathematical statement for "Aspirin has no causal effect on the headache outcome Y" is $Y^1 = Y^0$.
- Similarly, we can think of an individual with a beneficial causal effect of aspirin if $Y^1 > Y^0$, or one with a harmful 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.
- 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.

Counterfactual means

- Consider the following finite population version of the 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

- A commonly used population causal effect is given by the average causal effect (ACE)/ average treatment effect (ATE):

$$\psi = E[Y^1 - Y^0] = 1/4 - 1/2 = -1/4$$

- This estimand can be written as a functional of the two marginal distributions of Y^0 and Y^1, without requiring their joint distributions.

Identification through randomization

Suppose we randomize our population of patients with a headache to either aspirin or no aspirin with equal probability 1/2.

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

Based on the observed data,

$$E(Y|A=1) - E(Y|A=0) = 1/4 - 2/4 = -1/4$$

so that in this population, the crude association between A and Y appears to coincide with the average causal effect of A on Y.

Identification through randomization

(CA) Consistency Assumption: $Y = Y^A$ w.p. 1
(RA) Randomization Assumption: $(Y^0, Y^1) \perp\!\!\!\perp A$
(PA) Positivity Assumption: $0 < Pr(A = 1) < 1$
If these assumptions hold, then

$$\begin{aligned} \psi =& E(Y^1) - E(Y^0) \\ =& E(Y^1|A = 1) - E(Y^0|A = 0) \quad (RA) \\ =& E(Y|A = 1) - E(Y|A = 0) \quad (CA) \end{aligned}$$

Randomization

- Note that the randomization assumption is simply saying that since A is determined 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

Suppose that as often the case in randomized trials, we have also observed a vector L of pre-randomization risk factors for the outcome, then it is interesting to note that by randomization $L \perp\!\!\!\perp A$,

$$\begin{aligned} E(Y^a) =& E(Y|A=a) \\ =& \sum_l E(Y|A=a, L=l) f_L(l|a) \\ =& \sum_l E(Y|A=a, L=l) f_L(l) \end{aligned}$$

Therefore, the ATE has an alternative representation which incorporates baseline covariate information.

Observational study: a point exposure case

- Suppose that randomization no longer holds, because the observed data (L, A, Y) come from a point exposure observational study.

- L is a rich vector of covariates that satisfies:
  (NUCA) No unmeasured confounding assumption: $(Y^0, Y^1) \perp\!\!\!\perp A | L$
  There are no unmeasured confounders for the effect of A on Y.

- We also now require positivity: if $f_L(l) > 0$ then $Pr(A = a | L = l) > 0$

NUCA

- 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 mimics 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).

G-formula

- Under CA, NUCA and positivity, it is sufficient to identify $E(Y^a)$ and thus $\psi = E(Y^1) - E(Y^0)$.
  - For discrete L,

$$E\left(Y^a\right) = \sum_l E\left(Y|A = a, L = l\right) Pr\left(L = l\right)$$

  - For continuous L,

$$E\left(Y^a\right) = \int E\left(Y|A = a, L = l\right) f_L\left(l\right) \mathrm{d}l$$

G-formula

Without loss of generality, suppose L is categorical; then

$$\begin{aligned} E(Y^a) =& E(E(Y^a|L)) \\ =& \sum_l E(Y^a|L=l) f_L(l) \\ =& \sum_l E(Y^a|A=a, L=l) f_L(l) \quad (NUCA) \\ =& \sum_l E(Y|A=a, L=l) f_L(l) \quad (CA) \end{aligned}$$

G-formula

- The above formula is known as the direct standardization of $E(Y|A=a,L)$. It is a special case of the longitudinal g-formula.

- Under NUCA, we see that crude association is not causation, as $\sum_l E(Y|A=a,L=l)f_L(l) = E(Y^a) \neq E(Y|A=a) = \sum_l E(Y|A=a,L=l)f_L(l|A=a)$.

- However, if NUCA holds, and either of the following conditions holds:

$$Y \perp\!\!\!\perp L|A \quad \text{or} \quad A \perp\!\!\!\perp L$$

  then $E(Y^a) = E(Y|A=a)$ and L is not a confounder, so that this implies that RA actually holds.

  - In the former 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 latter case, $E(Y^a) = \sum_l E(Y|A=a,L=l)f_L(l) = \sum_l E(Y|A=a,L=l)f_L(l|A=a) = E(Y|A=a)$

G-formula

- Note that

$$E(Y^a) = \sum_{y,a^*,l} yf(y|A = a^*, L = l)1(a^* = a)f_L(l)$$

  has representation as expectation with respect of law that puts mass 1 at $A = a$.

- The observed data density at (y, a, l), $f(y|A = a, L = l)f(A = a|l)f_L(l)$ is transformed into one where A is degenerate at $A = a$, $f(y|A = a^*, L = l)1(a^* = a)f_L(l)$. The latter is often described as an intervention law.
- The ATE is completely invariant to the treatment process.

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 $E[Y^a] = g(a) = \sum_l E(Y|A = a, L = l)f_L(l)$.

- A natural nonparametric estimator of $g(a)$ is given by the nonparametric plug-in estimator, which requires nonparametric estimates of $E(Y|A = a, L = l) = b(a, l)$ and of $f_L(l)$.

G-computation

- Assume both A and L are categorical variables with low to moderate number of levels, so that $b(a, l)$ is given by the stratified sample average: $\hat{b}(a, l) = \sum_{i=1}^{n} I(A_i = a, L_i = l) Y_i / \sum_{i=1}^{n} I(A_i = a, L_i = l)$ and $\hat{f}_L(l) = 1/n \sum_{i=1}^{n} I(L_i = l)$.

- The nonparametric estimator of the G-formula is given by: $\hat{g}(a) = \sum_l \hat{b}(a, l) \hat{f}_L(l) = \sum_l \hat{b}(a, l) 1/n \sum_{i=1}^{n} I(L_i = l) = 1/n \sum_{i=1}^{n} \sum_l \hat{b}(a, l) I(L_i = l) = 1/n \sum_{i=1}^{n} \hat{b}(a, L_i)$

IPTW

- We discuss an alternative strategy for estimation of the g-formula, known as inverse probability treatment weighting (IPTW/IPW).

- The approach does not require estimating the outcome regression $Pr(Y = 1|A, L)$ or $E(Y|A, L)$ and instead relies on a model for the propensity score $Pr(A = 1|L)$, the probability of being exposed given the vector of confounders. This model may be well estimated even when Y is rare in the population because it does not require data on Y.

IPTW

Under the three assumptions, one can show that

$$E[Y^a] = E\left[\frac{I(A=a)}{f(A|L)}Y\right]$$

$$\begin{aligned}
E[\frac{I(A=a)}{f(A|L)}Y] =& E[\frac{I(A=a)}{f(a|L)}Y^a] \quad (CA) \\
=& E[\frac{E[I(A=a)|L, Y^a]}{f(a|L)}Y^a] \quad (IE) \\
=& E[\frac{f(a|L, Y^a)]}{f(a|L)}Y^a] \\
=& E[\frac{f(a|L)}{f(a|L)}Y^a] \quad (NUCA) \\
=& E(Y^a)
\end{aligned}$$

IPTW

$$E[Y^a] = E\left[\frac{I\left(A=a\right)}{f\left(A|L\right)}Y\right]$$

- The LHS is the average in the population of a latent counterfactual Y^a for all individuals in the population, while the RHS is a weighted average of the observed outcome Y for individuals with $A = a$, with weight $W = 1/f(a|L)$
- The weight W is the inverse probability of exposure given L. Therefore, $W > 1$ for all individuals contributing to the average, since $0 < f(a|L) < 1$.
- Each person counts for more than herself, and therefore weight creates a bigger pseudo-population than original observed subpopulation.
- This reweighting is carefully executed such that there is no confounding in the weighted sample.

IPTW

- For example, consider $E(Y^1)$, suppose that L is binary and $Pr(A=1|L=1)=1/4$, $Pr(A=1|L=0)=1/4$.
- This says that in the sample, on average 1 in every 4 people with $L=1$ received the intervention of interest $A=1$ and the other three did not follow the intervention of interest and therefore we do not observe their Y^1. The one person for whom Y^1 is observed is then upweigthed by 4 so that she counts not just for herself, but also for the other 3 people with $L=1$ who took $A=0$ and have missing Y^1.
- Weighting each person with $L=0$ is similar.

IPTW: an example

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 predicts Y given A.

IPTW: an example

- The crude mean

$$E(Y|A=1) = \sum_{l} E(Y|A=1, L=l) f(L=l|A=1)$$
$$= 24 \times 4/7 + 36 \times 3/7 = 204/7$$

- Whereas

$$E(Y^1) = \sum_{l} E(Y|A=1, L=l) f(L=l)$$
$$= 24 \times 1/2 + 36 \times 1/2 = 210/7$$

IPTW: an example

- Create a pseudo population by reweigthing the numbers in each row by $f_{A|L}^{-1}$

N	$f(A\|L)$	Pseudo$-N$	A	L	$E(Y\|A=a,L=l)$
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:

$$E_{ps}(Y|A=1) = 24 \times 1/2 + 36 \times 1/2 = 210/7 = E(Y^1)$$

Dual representation

We have described two alternative ways of estimating ATE.

- G-computation which requires a model for the outcome regression $E(Y|A, L)$.
- IPTW which requires a model for the propensity score $Pr(A = 1|L)$.

When the dimension of L is low, e.g., single binary L, then nonparametric estimation of the outcome regression and the propensity score produce the exact same point estimate and therefore, there is no advantage of one approach over the other.

Dual representation

$$\begin{aligned}
\hat{g}(a) =& \sum_l \hat{b}(a,l)\hat{f}_L(l) \\
=& 1/n \sum_l \sum_{i=1}^n \hat{b}(a,l) I(L_i = l) \\
=& 1/n \sum_l \sum_{i=1}^n I(L_i = l) \frac{\sum_{s=1}^n I(A_s = a, L_s = l) Y_s}{\sum_{j=1}^n I(A_j = a, L_j = l)} \\
=& 1/n \sum_l \sum_{s=1}^n I(A_s = a, L_s = l) Y_s \left[\frac{\sum_{i=1}^n I(L_i = l)}{\sum_{j=1}^n I(A_j = a, L_j = l)} \right] \\
=& 1/n \sum_{s=1}^n I(A_s = a) Y_s \hat{f}_{A|L}^{-1}(A_s|L_s)
\end{aligned}$$

Estimating equation perspective

- One can easily show that the IPTW estimator $\hat{g}(a)$ is the solution to the estimating equation:

$$\sum_{i=1}^{n} \frac{I(A_i = a)}{\hat{f}_{A|L}(a|L_i = l)}(Y_i - \hat{g}(a)) = 0.$$

- In contrast to the estimating equation for the conditional mean $\mu(a) = E(Y|A = a)$ given by

$$\sum_{i=1}^{n} I(A_i = a)(Y_i - \hat{\mu}(a)) = 0.$$

- 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 $\hat{f}_{A|L}^{-1}(a|L_i = l)$ to adjust for confounding by L.

IPTW in a randomized trial: an efficiency paradox

- We revisit the familiar setting where A is randomized. Denote by L a pretreatment categorical covariate that is a strong predictor of the outcome Y.
- By randomization, we have $(Y^0, Y^1, L) \perp\!\!\!\perp A$.
- As before, the crude estimator

$$\tilde{\psi} = \frac{\sum_{s=1}^n I(A_s = 1) Y_s}{\sum_{s=1}^n I(A_s = 1)} - \frac{\sum_{s=1}^n I(A_s = 0) Y_s}{\sum_{s=1}^n I(A_s = 0)}$$

is a valid estimator of the average causal effect ψ.

- Note that this estimator may also be written as an IPTW estimator with known weights $f_{A|L}^{-1}(A_s|L_s) = (1/2)^{-1}$

$$\tilde{\psi} = \frac{\sum_{s=1}^n I(A_s = 1) Y_s (1/2)^{-1}}{\sum_{s=1}^n I(A_s = 1)(1/2)^{-1}} - \frac{\sum_{s=1}^n I(A_s = 0) Y_s (1/2)^{-1}}{\sum_{s=1}^n I(A_s = 0)(1/2)^{-1}}$$

IPTW in a randomized trial: an efficiency paradox

- Compare this estimator to the one used by a statistician who decides to ignore randomization but rather assumes that $(Y^0, Y^1) \perp\!\!\!\perp A|L$, so that she may use G-computation by

$$
\begin{aligned}
\hat{\psi} =& \hat{g}(1) - \hat{g}(0) \\
=& \sum_l [\hat{b}(1, l) - \hat{b}(0, l)]\hat{f}_L(l) \\
=& \frac{\sum_{s=1}^n I(A_s = 1) Y_s \hat{f}_{A|L}^{-1}(A_s|L_s)}{\sum_{s=1}^n I(A_s = 1) \hat{f}_{A|L}^{-1}(A_s|L_s)} - \frac{\sum_{s=1}^n I(A_s = 0) Y_s \hat{f}_{A|L}^{-1}(A_s|L_s)}{\sum_{s=1}^n I(A_s = 0) \hat{f}_{A|L}^{-1}(A_s|L_s)}
\end{aligned}
$$

IPTW in a randomized trial: an efficiency paradox

- Compare this estimator to the one used by a statistician who decides to ignore randomization but rather assumes that $(Y^0, Y^1) \perp\!\!\!\perp A | L$, so that she may use G-computation by

$$
\begin{aligned}
\hat{\psi} =& \hat{g}(1) - \hat{g}(0) \\
=& \sum_l [\hat{b}(1,l) - \hat{b}(0,l)]\hat{f}_L(l) \\
=& \frac{\sum_{s=1}^n I(A_s = 1) Y_s \hat{f}_{A|L}^{-1}(A_s|L_s)}{\sum_{s=1}^n I(A_s = 1) \hat{f}_{A|L}^{-1}(A_s|L_s)} - \frac{\sum_{s=1}^n I(A_s = 0) Y_s \hat{f}_{A|L}^{-1}(A_s|L_s)}{\sum_{s=1}^n I(A_s = 0) \hat{f}_{A|L}^{-1}(A_s|L_s)}
\end{aligned}
$$

- $var(\hat{\psi}) \leq var(\tilde{\psi})$

IPTW in a randomized trial: an efficiency paradox

- But this is paradoxal as we find that the estimator $\tilde{\psi}$ that uses available information on the treatment mechanism, is less efficient than $\hat{\psi}$ by which ignores this information.
- The paradox is resolved by realizing that $\tilde{\psi}$ uses the available information, but in a very inefficient way as it ignores the fact that L is a strong correlate of Y.
- More information is generally a good thing, but not helpful if used inefficiently.

Other propensity score methods

- The propensity score has the following important property: if $\{Y^a : a \in supp(A)\} \perp\!\!\!\perp A | L$ then $\{Y^a : a \in supp(A)\} \perp\!\!\!\perp A | f_{A|L}(1|L)$.
- Therefore, to adjust for confounding by a large vector of confounders L, it is sufficient to adjust for the one dimensional summary $f_{A|L}(1|L)$.

Other propensity score methods

- Propensity score regression. Note that

$$\begin{aligned} E(Y^a) &= \int E(Y^a | f_{A|L}(1|L)) \mathrm{d}F(f_{A|L}(1|L)) \\ &= \int E(Y^a | A = a, f_{A|L}(1|L)) \mathrm{d}F(f_{A|L}(1|L)) \\ &= \int E(Y | A = a, f_{A|L}(1|L)) \mathrm{d}F(f_{A|L}(1|L)) \end{aligned}$$

  is a function of the observed data only. It is typically estimated in two stages:

  1. Regress A on L to obtain $\hat{f}_{A|L}(A|L;\hat{\alpha})$
  2. Regress Y on $\{A, \hat{f}_{A|L}(1|L;\hat{\alpha})\}$ to obtain $\hat{E}[Y|A = a, \hat{f}_{A|L}(1|L, \hat{\alpha})]$

Other propensity score methods

- Thus the approach requires both correct specification of $\hat{f}_{A|L}(A|L;\hat{\alpha})$, but also correct specification of $\hat{E}[Y|A, \hat{f}_{A|L}(A|L;\hat{\alpha})]$.

- This requires more modeling assumptions than IPTW, outcome regression, and DR estimators, respectively.

Other propensity score methods

Suppose that $E(Y|A=1,L)-E(Y|A=0,L)=\psi$, is constant within levels of L. Then the OLS of the working model

$$E[Y|A=a, f_{A|L}(1|L;\alpha);\eta] = \psi A + \eta_1 f_{A|L}(1|L;\alpha) + \eta_0$$

is fully robust, in the sense that

$$\hat{\psi} = \int \hat{E}[Y|A=1, \hat{f}_{A|L}(1|L;\hat{\alpha});\hat{\eta}]\mathrm{d}F(\hat{f}_{A|L}(1|L;\hat{\alpha})) \\ - \int \hat{E}[Y|A=0, \hat{f}_{A|L}(1|L;\hat{\alpha});\hat{\eta}]\mathrm{d}F(\hat{f}_{A|L}(1|L;\hat{\alpha}))$$

is consistent for $\psi = E(Y^1)-E(Y^0)$ even if $E[Y|A=a, f_{A|L}(1|L)] \neq E[Y|A=a, f_{A|L}(1|L);\eta]$ for all η.

Other propensity score methods

- The previous robustness result relies not only on the assumption of no exposure-confounder interaction, but also on the identity link.
- The robustness does not hold for nonlinear link: say if a working model

$$g(E(Y|A=a, f_{A|L}(1|L,\alpha);\eta)) = \psi A + \eta_1 f_{A|L}(1|L,\alpha) + \eta_0$$

  is mis-specified, where g is say the log or logit link, then neither ψ nor $\int E(Y|A=a, f_{A|L}(1|L,\alpha);\eta)\mathrm{d}F(f_{A|L}(1|L))$ has a causal interpretation, even if the effect of A is additive within levels of L on the g scale.

Other propensity score methods

- Matching.

$$
\begin{aligned}
A\hat{T}T &= \frac{1}{N_1} \sum_{\{i:A_i=1\}} [Y_{1i} - \hat{Y}_{0i}], \\
A\hat{T}NT &= \frac{1}{N_0} \sum_{\{i:A_i=0\}} [\hat{Y}_{1i} - Y_{0i}], \\
A\hat{T}E &= \frac{1}{N} \sum_{i=1}^{N} [\hat{Y}_{1i} - \hat{Y}_{0i}].
\end{aligned}
$$

Doubly robust estimator

$$
\begin{aligned}
EIF(\psi) =& \frac{A}{f(1|L)}\{Y - E(Y|1, L)\} + E(Y|1, L) \\
& - [\frac{1-A}{f(0|L)}\{Y - E(Y|0, L)\} + E(Y|0, L)] - \psi
\end{aligned}
$$

$$
\begin{aligned}
\hat{\psi}_{DR} =& \mathbb{P}_n \Big[\frac{A}{\hat{f}(1|L)}\{Y - \hat{E}(Y|1, L)\} + \hat{E}(Y|1, L) \\
& - [\frac{1-A}{\hat{f}(0|L)}\{Y - \hat{E}(Y|0, L)\} + \hat{E}(Y|0, L)] \Big]
\end{aligned}
$$

Doubly robust estimator

A simple construction of DR estimator:

- Estimate α using

$$f_{A|L}\left(1 \mid L;\alpha\right)=\left(1+\exp\left(-L'\alpha\right)\right)^{-1}.$$

- Consider Y continuous, specify $b(A,L;\eta)=(A,L')\,\eta$, and obtain OLS of η.
- Obtain $\hat{\psi}_{DR}$.
- Then one can show that $\hat{\psi}_{DR}$ is consistent if either $E(Y \mid A,L)$, or $f_{A|L}\left(A \mid L\right)$ but not necessarily both are correctly specified.

Product bias property

$$bias(\psi_{DR}) = \sum_a (-1)^{1-a} \left(\frac{f(a|L)}{\tilde{f}(a|L;\alpha)} - 1 \right) \left(E(Y|a,L) - \tilde{E}(Y|a,L;\eta) \right)$$

G-estimation

- Let $\gamma_0(a, l)$ encodes the individual causal effect

$$\gamma_0(A, L) = Y^A - Y^0 \ wp\ 1$$

- Note that $\gamma_0(a, l) = 0$ encodes the sharp null hypothesis that $Y^a = Y^0$ for all a.
- With consistency, we have

$$Y^0 = Y - \gamma_0(A, L) \ wp\ 1$$

- By NUCA, we have

$$Y - \gamma_0(A, L) \perp\!\!\!\perp A | L$$

G-estimation

- We consider a model $\gamma(A, L; \psi)$ such that $\gamma(A, L; \psi = 0) = \gamma(A = 0, L; \psi) = 0$ and $\gamma(A, L; \psi_0) = \gamma_0(A, L)$. For example, for binary A one might consider
  (1) $\gamma(A, L; \psi) = A\psi$; which assumes no effect heterogeneity.
  (2) $\gamma(A, L; \psi) = (A, AL')\psi$; which models effect heterogeneity as a linear function of L.
- Thus our goal is to make inferences about the finite dimensional parameter ψ_0. Suppose that A is binary and that the propensity score model $\pi(L) = Pr(A = 1|L)$ is known. Then, let

$$H(\psi) = Y - \gamma(A, L; \psi)$$

- The restriction motivates the following estimating equation for ψ_0

$$E[H(\psi_0)(A - \pi(L))m(L)] = 0$$

for m a user-specified function of same dimension as ψ.

G-estimation

- In practice, $\pi(L)$ is not known and must be estimated from data on A and L; let $\hat{\pi}(L) = \hat{\pi}(L; \hat{\alpha})$ denote such an estimator based, for instance, on logistic regression.
- G-estimation solves

$$\mathbb{P}_n[H(\hat{\psi}_g)(A - \hat{\pi}(L))m(L)] = 0$$

- For instance, consider model (1). Then one can choose $m = 1$ and solve for

$$\hat{\psi}_g = \frac{\mathbb{P}_n[Y(A - \hat{\pi}(L))]}{\mathbb{P}_n[A(A - \hat{\pi}(L))]}$$

G-estimation

- G-estimation is a semiparametric approach in the sense that it allows for estimation of ψ_0 without modeling the outcome distribution under a correctly specified model for the propensity score, denoted by $\mathcal{M}_\pi$.

- Note that under $\mathcal{M}_\pi$ one is free to augment the estimating equation as follows

$$\mathbb{P}_n[H(\hat{\psi}_g)(A-\hat{\pi}(L))m(L)]+\mathbb{P}_n[t(L)(A-\hat{\pi}(L))]=0$$

for any function $t(L)$.

Doubly robust G-estimation

- For fixed $m(L)$, the optimal choice of t in model $\mathcal{M}_\pi$ is given by

$$t(L) = -E[H(\psi_0)|L]m(L)$$

- Note that by NUCA and consistency, $E[H(\psi_0)|L] = E(Y|A=0, L)$. As $E[H(\psi_0)|L]$ is unknown, let $\hat{E}(H(\psi)|L)$ denote an estimator of $E(H(\psi)|L)$ for fixed ψ using a parametric model, e.g., $E(H(\psi)|L;\eta) = (1, L')\eta$, and $\hat{\eta}$ solves the equations

$$\mathbb{P}_n[(1, L')'(H(\psi) - (1, L')\eta)] = 0$$

- This motivates the following estimating equation

$$\mathbb{P}_n[\{H(\hat{\psi}_{DR}) - E(H(\hat{\psi}_{DR})|L;\hat{\eta})\}(A - \hat{\pi}(L;\hat{\alpha}))m(L)] = 0$$

Outcome regression

- It is interesting to note that under model $\mathcal{M}_\eta$ a standard approach for estimating ψ_0 in fact entail performing the OLS $(H(\psi) = Y - \psi A)$

$$\mathbb{P}_n \left[\begin{pmatrix} 1 \\ L \\ A \end{pmatrix} (Y - \hat{\psi}_{OR} A - (1, L')\hat{\eta}) \right] = 0$$

- This approach relies entirely on the assumption that $Y = \psi_0 A + (1, L')\eta + \epsilon$ where $E(\epsilon | A, L) = 0$. This is sometimes referred to as outcome regression estimation.

G-estimation

- The rank preserving model $Y^0 = Y^A - \gamma_0(A, L)$ *wp* 1 is unrealistic in the health sciences, as it rules out the presence of unobserved predictors of Y that interact with treatment. Biologically, we might expect for instance that there would be some interactions with unmeasured genetic factors that are unknown to the analyst.

- Fortunately, the estimating equation approach described previously continues to hold, if the model is defined on the population mean scale:

$$\gamma_0(a, l) = E(Y^a - Y^0 | L = l)$$

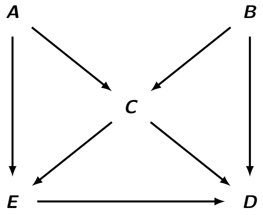
  so that $\gamma_0(a, l)$ captures the difference of counterfactual means were one to intervene and set treatment to $A = a$ vs. $A = 0$ among the subpopulation with $L = l$.

Causal diagrams

- Consider a hypothetical study of the relation of antihistamine treatment to asthma incidence 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. There are no other important confounders besides sex, bronchial reactivity and air pollution

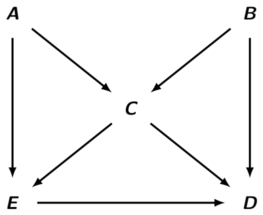
Causal diagrams

These assertions can be incorporated in the following diagram



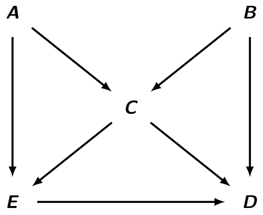
- The points representing the variables are called the vertices/nodes of the graph; any line or arrow connecting two variables in the graph is an arc/edge.
- Arrows represent direct links from causes to effects, that is not mediated by any other variable. Example: the arrow linking A and C represents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

Causal diagrams



- Absence of arrow ⇔ no direct causal effect. For example, no arrow from A to D reflects assertion (3).
- A node within a path is said to intercept the path. For example, C intercepts the paths A-C-D and E-C-D.

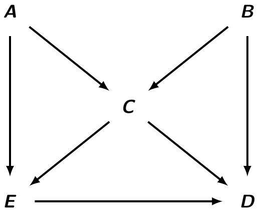
Causal diagrams



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

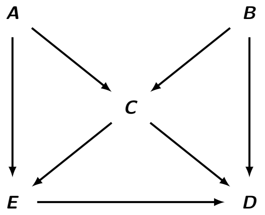
- X is a parent of Y if there is a single headed arrow from X into Y: in such a case Y is called a child of X. For example, A and C are parents of E, whereas C and E are children of A.

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. For example, all paths from E to D except the direct path.
- A path collides at a variable X if the path enters and exits 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. For example,
    - 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.

Directed Acyclic Graph (DAG)



- The diagram for this study is a Directed Acyclic Graph (DAG)
  - Directed: all arcs between variable are arrows. Moreover, directed path ⇔ causal path.
  - Acyclic if no directed path in the graph forms 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.
  - One does not need to include variables not of interest or not common causes of variables in the DAG.

DAGs: remarks

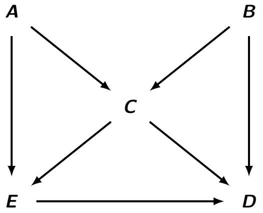
- A DAG is nonparametric, does not impose any functional restriction on the joint distribution of the variables on the graph.
- Production of an effect by a cause requires a directed path from the cause to the effect on the graph, e.g., absence of a directed path between A and B implies absence of a causal effect between the two variables, which also implies there is no effect of A on D through B.
- Graphs also encode associations between variables: absence of an unblocked path between two variables implies statistical independence of 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.
- In fact there are only two reasons two variables are statistically dependent:

  * They share a common cause, or

  * One variable causes the other.

DAGs: remarks

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



- In our causal DAG, observed data pdf factorizes as

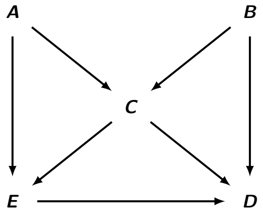
$$f(A, E, C, B, D) = f(D|B, C, E)f(C|A, B)f(E|A, C)f(A)f(B)$$

where we use the Markov factorization, that a variable is independent of nonparental ancestors given its parents:

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

DAGs: remarks

- 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.



DAGs: remarks

- 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.
- 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

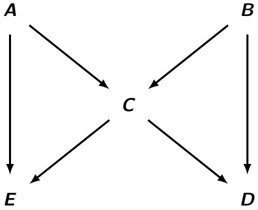
- Confounding occurs when the study exposure groups differ in their probability distribution for the outcome for reasons other than exposure effect.
- Such differences are attributable to effects of extraneous variables which are called confounders.
- Counfounding is present if and only if exposure would 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:
  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.

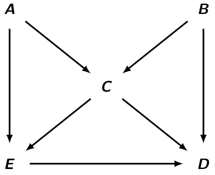
DAGs and confounding

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



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

DAGs and confounding



A conventional approach is to condition on potential confounders:

- 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-D, 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 be 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

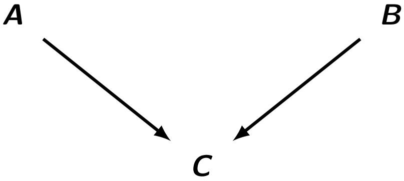
	$A = 1$		$A = 0$	
	$B = 1$	$B = 0$	$B = 1$	$B = 0$
$C = 1$	800	600	400	200
$C = 0$	200	400	600	800
Total	1000	1000	1000	1000

DAGs and confounding

We have

$$Pr(A = 1|B) = Pr(A = 1) = Pr(B = 1|A) = Pr(B = 1) = 0.5.$$

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

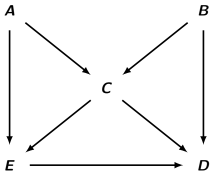


DAGs and confounding

The conditional odds ratio for the A-B association is not one 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.

DAGs and confounding



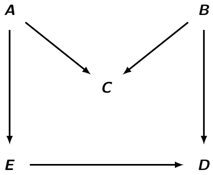
Returning to our original DAG, applying this rule tells us that conditioning on C can create an unblocked back door path from E to D: the association of A and B within levels of C can create an association of A and D indirectly, through B.

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.

DAGs and confounding

Another example where things can go terribly wrong:



Clearly there is 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.

d-separation

- Definition: A trail between nodes A and Y in a DAG G is said to be d-separated or blocked by a set of nodes B disjoint with $\{A, Y\}$ iff
  a The trail contains a collider and neither the collider nor any of its descendants are in B; or
  b The trail contains a chain or a fork whose vertex is in B.

- Definition: Node A and node Y are said to be d-separated or blocked by a set of nodes B in DAG G if all trails between A and Y in G are d-separated by B.

d-separation

- Definition: If one or both A and Y are comprised by a set of nodes, instead of just a single node, and A, Y and B are pairwise disjoint, then A and Y are said to be d-separated or blocked by B in G iff B d-separates in G every node in A from every node in Y.

- It is therefore clear that B d-separates two sets of variables A and Y if the following hold:
  * Every unblocked path from A to Y is intercepted by a variable in B.
  * Every unblocked path from A to Y generated by adjustment for the variables in B is intercepted by a variable in B.

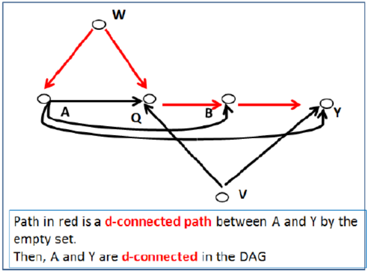
d-separation

- Nodes that are not d-separated are said to be unblocked or d-connected.

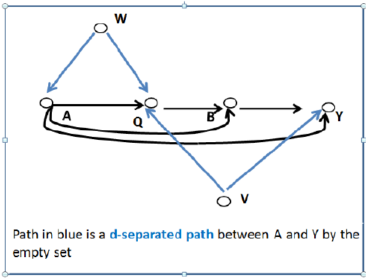
- The letter d in "d-separation" stands for "separation in the DAG".

- Note that d-separation of A and Y by B in DAG G sometimes is denoted as $(A \perp\!\!\!\perp Y | B)_G$. This should not be confused with $A \perp\!\!\!\perp Y | B$ which stands for A and Y are conditionally independent given B.

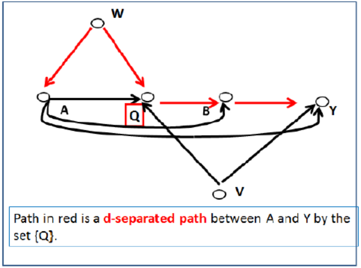
d-separation



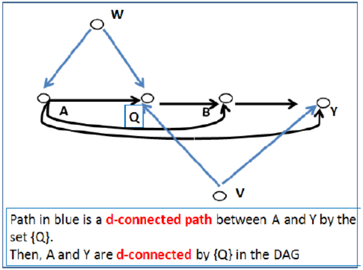
Path in red is a d-connected path between A and Y by the empty set.
Then, A and Y are d-connected in the DAG



Path in blue is a d-separated path between A and Y by the empty set



Path in red is a d-separated path between A and Y by the set {Q}.



Path in blue is a d-connected path between A and Y by the set {Q}.
Then, A and Y are d-connected by {Q} in the DAG

Back door path criterion

A set of variables S is sufficient for control of confounding under a given DAG if S contains no descendants of E, and S d-separates E from D in the graph obtained by deleting all arrows emanating from E. In the asthma example, back door criterion gives $S = \{A, C\}$ or $S = \{B, C\}$.