Introduction to Causal Inference

Erin E Gabriel

SMARTbiomed
Pioneer Centre

Outline

Outline

Let's talk about bias

We have an observational cohort collected in 2021 where some patients take COVID-19 vaccine and some do not. We run an analysis for the side-effect of liver stiffness.

- Y liver stiffness
- A vaccine exposure

$$\theta = P\{Y = 1|A = 1\} - P\{Y = 1|A = 0\}$$

Causal interpretation?

- Can we give θ a causal interpretation? If so, what exactly is this causal interpretation?
- As in standard statistics, we want to consider the asymptotic bias of our estimator of the log odds ratio, but bias in terms of what?

$$\theta - \{\textit{true-value-of-some-causally-interpretable-estimand}\}$$

- Under what assumptions is the above asymptotic bias zero?
- Without the mathematical vocabulary even to write down what we're interested in, there is very little hope of rigorous progress...

Randomization

Instead we randomize people to take COVID-19 vaccine or not. We run an analysis for the side-effect of liver stiffness.

- Y liver stiffness
- A vaccine exposure

$$P\{Y = 1|A = 1\} - P\{Y = 1|A = 0\}$$

If you see that there is a increase in liver stiffness, is this causal? **What is the difference?**

Marginal causal effect

For a binary outcome, Y and a binary exposure A

- marginal causal risk difference, for the full population

$$p(Y^1 = 1) - p(Y^0 = 1)$$

- marginal causal risk ratio, for the full population

$$\frac{p(Y^1 = 1)}{p(Y^0 = 1)}$$

- marginal causal odds ratio, for the full population

$$\frac{p(Y^1 = 1)}{p(Y^1 = 0)} \Big/ \frac{p(Y^0 = 1)}{p(Y^0 = 0)}$$

Conditional causal effect

V a set of variables

- Conditional causal risk difference, given $V = v$

$$p(Y^1 = 1|v) - p(Y^0 = 1|v)$$

- Conditional causal risk ratio, given $V = v$

$$\frac{p(Y^1 1 = 1|v)}{p(Y^0 = 1|v)}$$

- Conditional causal odds ratio, given $V = v$

$$\frac{p(Y^1 = 1|v)}{p(Y^1 = 0|v)} \Big/ \frac{p(Y^0 = 1|v)}{p(Y^0 = 0|v)}$$

A fundamental problem of causation

Only one (max) of the counterfactual or potential outcomes can be observed at any time for any subject. Thus, we cannot estimate directly $Y_i^1 - Y_i^0$ without wild additional assumptions.

For this reason, estimands such as $P(Y_i^1 > Y_i^0)$, the probability of benefit, and $E\{A|Y_i^1 = Y_i^0\}$, the expected value of x within the principal stratum where $Y^1 = Y^0$, are controversial.

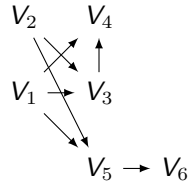
Outline

Non-causal Directed acyclic graph

- A directed acyclic graph (DAG) is a way of representing the factorization of a joint distribution.
- For example, if $\mathbf{V} = (V_1, V_2, V_3, V_4, V_5, V_6)$ and supposing the joint distribution factorizes as:

$$p_{\mathbf{V}}(\mathbf{v}) = p_{V_1}(v_1)p_{V_2}(v_2)p_{V_3|V_1,V_2}(v_3|v_1, v_2) \cdot p_{V_4|V_1,V_3}(v_4|v_1, v_3)p_{V_5|V_1,V_2}(v_5|v_1, v_2)p_{V_6|V_5}(v_6|v_5)$$

then we can represent this factorization by the following graph:

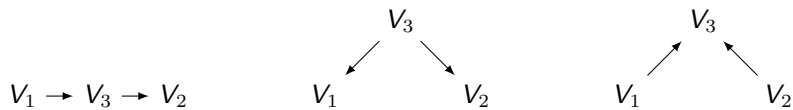


Pathways

Pathways give you information about associations.

- direct pathways between two variables show those variables are directly associated, i.e. you cannot factor the joint so that the variable at the end of the arrow does not depend on the variable at the start of the arrow.
- pathways between two variables, even not direct ones imply associations, but the arrow direction matters.

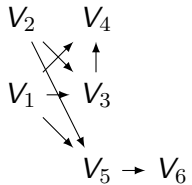
Summary of three-variable examples



- V_1 and V_2 are (marginally) associated in situations 1 and 2, but (marginally) independent in situation 3, i.e. not conditioning on any other variables the joint cannot be factorized such that the distribution of V_2 does not depend on V_1 in the first two figures.
- Conditioning on V_3 removes the association (between V_1 and V_2) in situations 1 and 2. V_3 blocks the pathway in 1 and 2.
- Conditioning on V_3 (a collider) creates an association (between V_1 and V_2) in situation 3.

d-separation

L d-separates A from Y in the graph if and only if L blocks every path from A to Y.



Does V_2 d-separate V_3 from V_6 in the graph above? You have ten minutes to work with out. Please try.

NO

- Here are the paths from V_3 to V_6:
  $V_3 \leftarrow V_1 \rightarrow V_5 \rightarrow V_6$
  $V_3 \leftarrow V_2 \rightarrow V_5 \rightarrow V_6$
  $V_3 \rightarrow V_4 \leftarrow V_1 \rightarrow V_5 \rightarrow V_6$
  The first path is not blocked by V_2. The second path is blocked by V_2 and in the third path, V_4 is a collider that is not open. Therefore V_2 does not d-separate V_3 from V_6.

Causal DAGs

$$A \longrightarrow Y \qquad A \longleftarrow Y$$

- As representations of joint distributions, these 2 DAGs are equivalent.
- However, as causal DAGs, their interpretation is different.
- The first DAG, read as a causal DAG, says that A (possibly) affects Y, but that Y definitely does not affect A.
- The second, read as a causal DAG, says that Y (possibly) affects A, but that A definitely does not affect Y.
- The 'possibly' statement could be dropped: it is the absence of arrows that represent assumptions in DAGs.

Rules for Causal DAGs

- A causal DAG is a special sort of DAG, so it too must correspond to a factorization of the joint distribution.
- But it can include unobserved variables; indeed, usually it MUST because:
- Any common cause of two variables in the causal DAG must itself be in the causal DAG. A cause of just one variable need not be included.
- **acyclic** means no variable can cause itself, even via other variables. Thus, no cycles.

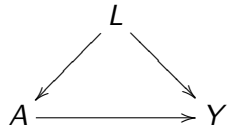
Meanings that everyone agrees with

- A causal DAG illustrates some of our beliefs about how the data were generated, specifically the absence of a direct arrow from A to B means A does not have a direct causal effect on B (relative to the other variables in the causal DAG).
- Generally, the data can't tell us exactly which causal DAG to draw; this must be partly based on subject-matter knowledge, But certain causal DAGs can be ruled out by the data (if they imply conditional independencies that are violated). Causal discovery is a set of methods for discoverying a DAG (or some other structure) from the data.

NOT everyone agrees with DAGs or the encoded NPSEM

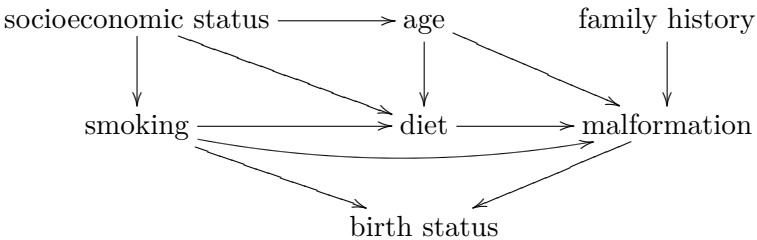
- Nonparametric structural equation models being directly encoded from the DAG is a take some people have Pearl [2000]
- Questionable usefulness for solving subject-matter problems Aronow and Sävje [2020]
- Some causal assumptions cannot be easily encoded in a DAG, e.g., monotonicity, functional forms of relations Aronow and Sävje [2020]
- others believe that it is just a representation of causal independencies, but not the full picture, and in particular, they don't believe in the cross-world independence assumptions, which we will get to with mediation

Underlying assumptions and pathways



- Assumptions are encoded by the direction of arrows
  - the arrow from A to Y means that A may affect Y, but not the other way around
  - if we are interested in knowing if there is a direct path from A to Y we need to block the path via L

A possible DAG in a real example



Graphical approach: Ideal randomized trials***

- In ideal randomized trials we have exchangeability:

$$(Y^0, Y^1) \perp\!\!\!\perp A$$

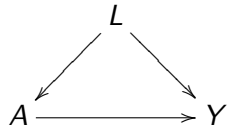
- As a consequence, association $=$ causation

And this is the DAG

$$A \longrightarrow Y$$

This trial probably doesn't exist.

Graphical approach: Observational studies



- In observational studies we usually don't have exchangeability, because of confounding
- We may achieve conditional exchangeability by controlling for an appropriate set of covariates:

$$(Y(0), Y(1)) \perp\!\!\!\perp A \mid L$$

Outline

Before we can estimate

We need to determine if the estimand is estimable, i.e., identifiable, and under what assumptions.

Note I will now use the $Y(a)$ notation for potential outcomes.

Causal estimands

- Using a causal language (such as potential outcomes) we can express causal estimands, e.g. for a binary outcome:

$$\mathsf{P}\left(Y^1 = 1\right) - \mathsf{P}\left(Y^0 = 1\right)$$

What does identification mean?

It means you can write the estimand without the counterfactual in terms of things you can observe and, therefore, estimate given your data.

This is, therefore, contextual: estimands are identifiable (or not) in a given settings and under a given set of assumptions. This may be non-parametrically or parametrically identified. unless otherwise noted people mean non-parametrically when speaking about identification

Non-identifiable sometimes refers to an estimand that someone deems to be never identified, often because they find the assumptions need to be objectionable or crazy. these are often used interchangeably in the literature.

In an ideal randomized trial, the DAG might look like this:

$$A \rightarrow Y$$

- Thus, (exchangeable) or (no unmeasured confounders)

$$(Y^0, Y^1) \perp\!\!\!\perp A$$

- Thus, $E\{Y^a\} = E\{Y|A = a\}$ $\underline{\text{IF}}$ $Y^a = Y$ if $A = a$

In the ideal randomized trial, the ATE is identified under the above two assumptions (which is actually three assumptions).

No interference

- Y^a represents the value that Y would have taken for individual i had A been set to a for individual i, i.e. the potential value of Y_i had A_i been set to a ($i = 1, \ldots, n$).
- By writing it this way we are already making the assumption that the potential value of Y_i does not depend on what A_j was set to ($j \neq i$).
- This assumption is called no interference.
- It can be relaxed, but definitions/assumptions/methods all get more complicated.
- An example where interference is present is in the study of political views, where influential individual j may affect the opinion of individual i.
- A different type of interference in contagion in infectious disease

Interference topic for presentation

Consistency (1)

- The second assumption we make is consistency:

$$Y^a = Y \text{ if } A = a$$

- For those who actually received exposure level a, their observed outcome is the same as it would be if they received exposure level a via the hypothetical intervention we have in mind.
- So, it helps to be clear about the precise nature of the intervention (or interventions) we have in mind, and for consistency to hold, having $A = a$ by such an intervention must lead to the same Y as if A turns out to be a in the observational setting.

Conditional exchangeability

- The third assumption we make is conditional exchangeability given some set of observed covariates $\mathbf{L}$:

$$Y^a \perp\!\!\!\perp A \,|\mathbf{L}\,, \forall a$$

- Conditional on $\mathbf{L}$, the actual exposure level A is independent of each of the potential outcomes.
- We can think of $\{Y^a, \forall a\}$ as capturing those characteristics of an individual relevant for Y except for A.
- Note that *if* we knew $\{Y^a, \forall a\}$ and we know A, then (under consistency) we would know Y.
- So conditional exchangeability says that these other relevant characteristics (other than A, that determine Y), represented by $\{Y^a, \forall a\}$, must be independent of A given $\mathbf{L}$.
- This is a rigorous formulation of the intuitive idea of 'no unmeasured confounders/confounding'.

Testability of exchangeability

Exchangeability, no unmeasured confounders, is not testable. You have to assume it, but there are sensitivity analyses and other ways around this assumption. Coming later today!

Identification (1)

- Suppose A and Y are both binary and we are interested in the marginal causal risk difference:

$$\Pr\left(Y^1 = 1\right) - \Pr\left(Y^0 = 1\right)$$

  [We are assuming 'no interference' throughout.]

- For simplicity, suppose that conditional exchangeability holds given one single discrete covariate L.
- By the law of total probability:

$$\mathrm{P}\left(Y^a = 1\right) = \sum_c \mathrm{P}\left(Y^a = 1 \,|L = l\right) \Pr\left(L = l\right)$$

- By conditional exchangeability, this can be rewritten as:

$$\sum_c \mathrm{P}\left(Y^a = 1 \,|A = a, L = l\right) \mathrm{P}\left(L = l\right)$$

Identification (2)

- By consistency, this is:

$$\sum_l \Pr\left(Y = 1 \,|A = a, L = l\right) \mathrm{P}\left(L = l\right)$$

- Thus we have:

$$\begin{aligned}
&\mathrm{P}\left(Y^1 = 1\right) - \mathrm{P}\left(Y^0 = 1\right) \\
&\qquad = \sum_l \mathrm{P}\left(Y = 1 \,|A = 1, L = l\right) \Pr\left(L = l\right) \\
&\qquad\qquad - \sum_l \mathrm{P}\left(Y = 1 \,|X = 0, L = l\right) \Pr\left(L = l\right)
\end{aligned}$$

- Note that we have, under the assumptions of conditional exchangeability (given L) and consistency, rewritten the causal estimand in terms of aspects of the distribution of the observed data: this is what is meant by *identification*. You can go from counterfactual $\rightarrow$ factual

Positivity

- At least in principle, given enough data, the estimation that follows identification can be non-parametric.

$$\mathrm{P}\left(Y = 1 \,|A = a, L = l\right)$$

for each a, l and

$$\mathrm{P}\left(L = l\right)$$

for each l can be replaced by observed proportions in our data.

- For this to work, there must be both exposed and unexposed individuals in each observed category of the confounder.
- This requires the positivity assumption:

Positivity: If $\Pr(L = l) > 0$ then: $0 < \Pr(A = 1|L = l) < 1$

Identifiable (according to Pearl)

The causal effect of A on Y is identifiable from a graph G if the quantity $P(Y|do(a))$ can be computed uniquely from any positive probability of the observed variables.

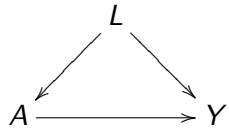
Thus we need to pieces to identify or estimate uniquely an estimand of interest:

- data
- assumptions/information about how that data was generated, either from the DAG or otherwise

One might then say that an estimand is identifiable given an assumed DAG allows for unique estimation via data generated under the DAG.

Graphs and NPSEM

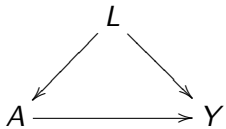
If we have the following DAG:



According to Pearl, this encodes nonparametric structural equations models (NPSEM):

$$
\begin{aligned}
l &= f_L(\epsilon_z) \\
a &= f_A(z, \epsilon_a) \\
y &= f_Y(x, z, \epsilon_y)
\end{aligned}
$$

Back-door criteria



- No node in L is a descendant of A; and
- L blocks every path between A and Y that contains an arrow into A (a back-door path)

Back-Door Adjustment formula

If L is measured, then conditioning on it blocks all back-door paths from A to Y, and thus, with the addition of consistency, all effects of A on Y are identified that are based on the marginal $E\{Y(a)\}$, via the back-door adjustment formula:

$$\sum_l P(y|a,l)P(l)$$

$$\int_l P(y|a,l)P(l)dl$$

Outline

Estimation

- Let us return to the back-door formula
- Recall that under conditional exchangeability and consistency:

$$\begin{aligned}\mathrm{P}\left\{Y^1 = 1\right\} - \mathrm{P}\left\{Y^0 =\right\} &= \int_{\mathbf{l}} \mathrm{P}\left(Y = 1 \,|A = 1, \mathbf{L} = \mathbf{l}\right) p_{\mathbf{L}}(\mathbf{l})\, d\mu_{\mathbf{L}}(\mathbf{l}) \\ &\quad - \int_{\mathbf{l}} \mathrm{P}\left(Y = 1 \,|A = 0, \mathbf{L} = \mathbf{l}\right) p_{\mathbf{L}}(\mathbf{l})\, d\mu_{\mathbf{L}}(\mathbf{l})\end{aligned}$$

- We could estimate

$$\mathrm{P}\left(Y = 1 \,|A = a, \mathbf{L} = \mathbf{l}\right)$$

  from a regression model as we have
- Integration over $\mathbf{L}$ is nonparametric (model-free).

The standardization formula

This is also referred to as the standardization formula

- If we have conditional exchangeability, given L is discrete, then

$$
\begin{aligned}
p(Y^0 = 1) &= \sum_L p(Y = 1 | A = 0, L) p(L) \\
p(Y^1 = 1) &= \sum_L p(Y = 1 | A = 1, L) p(L)
\end{aligned}
$$

Marginal effect vs. conditional effects

- In our example we observed that

$$p(L=0) = p(L=1) = 0.5$$

and

$$\begin{aligned} \frac{p(Y^1=1|L=0)}{p(Y^0=1|L=0)} &= 1 \\ \frac{p(Y^1=1|L=1)}{p(Y^0=1|L=1)} &= 2 \\ \frac{p(Y^1=1)}{p(Y^0=1)} &= 1.8 \end{aligned}$$

- The marginal effect is generally not equal to the average of the conditional effects, even if all levels of L are equally probable, because there are normally interactions.
- Exception: the causal risk (mean) difference often referred to as the average treatment effect ATE or the average causal effect ACE.

Marginal effect vs. conditional effects, cont'd

- The marginal effect is generally not equal to the average of the conditional effects, even if these are constant across levels of L
  - e.g. the conditional causal odds ratio may be equal to 3 for both males ($L = 1$) and females ($L = 0$),
  - but the marginal causal odds ratio may be equal to 1.5
- This is often referred to as 'non-collapsibility'
- Exceptions: the causal risk (mean) difference and the causal risk ratio (collapsible)

The logistic regression model

- Since the outcome is binary, it is natural to use the logistic regression model

$$\text{logit}\{p(Y = 1|A, L)\} = \alpha + \beta A + \gamma L$$

- What are the interpretations of α, β, and γ?

Solution

$$\text{logit}\{p(Y=1|A,L)\} = \alpha + \beta A + \gamma L$$

$$\begin{aligned}\alpha &= \text{logit}\{p(Y=1|A=0,L=0)\} \\ &= \log\left\{\frac{p(Y=1|A=0,L=0)}{p(Y=0|A=0,L=0)}\right\}\end{aligned}$$

$$\begin{aligned}\beta &= \text{logit}\{p(Y=1|A=1,L)\} - \text{logit}\{p(Y=1|A=0,L)\} \\ &= \log\left\{\frac{p(Y=1|A=1,L)}{p(Y=0|A=1,L)} \Big/ \frac{p(Y=1|A=0,L)}{p(Y=0|A=0,L)}\right\}\end{aligned}$$

$$\begin{aligned}\gamma &= \text{logit}\{p(Y=1|A,L+1)\} - \text{logit}\{p(Y=1|A,L)\} \\ &= \log\left\{\frac{p(Y=1|A,L+1)}{p(Y=0|A,L+1)} \Big/ \frac{p(Y=1|A,L)}{p(Y=0|A,L)}\right\}\end{aligned}$$

Causal interpretation

$$\text{logit}\{p(Y = 1|A, L)\} = \alpha + \beta A + \gamma L$$

- If we have conditional exchangeability, and the model is correctly specified, given L, then β is the conditional causal log odds ratio, given L

$$\beta = \log \left\{ \frac{p(Y^1 = 1|L)}{p(Y^1 = 0|L)} \Big/ \frac{p(Y^0 = 1|L)}{p(Y^0 = 0|L)} \right\}$$

Models and assumptions

Regardless of the target estimand, the model must be correctly specified for confounding, which means the model is statistically correctly specified, and L is sufficient for confounding control, i.e. removes all back-door paths from A to Y, leaving only the direct path (or the path of interest in the case of a mediator we are not conditioning on)

- **All models are wrong**
  - but if the model is approximately correct, then our conclusions are approximately valid
- Assumptions that we make should ideally be justified by both
  - subjects matter knowledge, and
  - diagnostic tests with data

The marginal effect

$$p(Y^0 = 1) \text{ vs } p(Y^1 = 1)$$

- Arguably more intuitive than main effect + interaction term
- Can always be presented as one single number (e.g. one log odds ratio) regardless of the number of interactions
- Informative about the 'general' treatment effect, since it applies to the whole population

The standardization formula

- If we have conditional exchangeability, given L, then

$$
\begin{aligned}
p(Y^0 = 1) &= \sum_L p(Y = 1|A = 0, L)p(L) \\
p(Y^1 = 1) &= \sum_L p(Y = 1|A = 1, L)p(L)
\end{aligned}
$$

- If L is binary, then we can estimate $p(Y = 1|A = 0, L)$ and $p(Y = 1|A = 1, L)$ without modeling assumptions
- If L is continuous, this non-parametric approach is not feasible
  - very few subjects for each observed level of L
- But with a regression model we can 'extrapolate', to estimate both $p(Y = 1|A = 0, L)$ and $p(Y = 1|A = 1, L)$

Four steps for standardization with regression model

- **Step 1**: fit a regression model for the outcome
- **Step 2**: replace the factual level of A with 0 for each subject
- **Step 3**: use the fitted model to estimate $p(Y = 1|A, L)$ for each subject, i.e. for each level of (A, L)
- **Step 4**: average these estimates to obtain an estimate of $p(Y^0 = 1)$
- To estimate $p(Y^1 = 1)$, replace A with 1 in step 2

Other measures of marginal effects

$$\begin{aligned}\hat{p}(Y^0 = 1) &= 0.6090828 \\ \hat{p}(Y^1 = 1) &= 0.06333356\end{aligned}$$

- Once we have estimated $p(Y^1 = 1)$ and $p(Y^0 = 1)$ we can estimate any measure of marginal causal effect, e.g.

$$\begin{aligned}\text{causal risk difference} &= \hat{p}(Y^1 = 1) - \hat{p}(Y^0 = 1) = -0.55 \\ \text{causal risk ratio} &= \hat{p}(Y^1 = 1)/\hat{p}(Y^0 = 1) = 0.10\end{aligned}$$

  even though the estimates were derived from a logistic regression model

Outline

Linear Regression for a continuous outcome?

When using back-door identification and linear regression for the continuous outcome and a binary exposure $X \in \{0, 1\}$

$$E\{(Y|A, L)\} = \alpha + \beta A + \gamma L$$

- Then the β coefficient from the model is a consistent estimate of the ATE if the model is correctly specified for confounding.

Linear Regression for a continuous outcome?

If you have an interaction in the linear regression model,

$$E\{(Y|A,L)\} = \alpha + \beta A + \gamma L + \nu L * A$$

Then β and ν may have causal interpretations, if the model is correctly specified for confounding, but neither is the marginal causal effect. standardization can again be used here.

Regression approach: why worry?

- Assumptions about the correct statistical specification of

$$E(Y|A,\mathbf{L}) = \alpha + \beta A + \boldsymbol{\gamma}^T \mathbf{L}$$

  can at least be checked from the data (unlike exchangeability), and adapted accordingly.

- We can add quadratic terms, product terms (interactions) etc. if the data suggest them.

- However:

  (1) For high-dimensional $\mathbf{L}$ (and we want $\mathbf{L}$ to be so, to adjust for as much confounding as possible!), this gets less feasible;
  (2) If there is little overlap in the values of some Ls between the $A = 0$ and $A = 1$ groups, we run into extrapolation problems (to be discussed next);
  (3) For high-dimensional $\mathbf{L}$ in non-linear models, the ML estimator of the exposure effect may suffer from finite sample bias

Propensity scores

The propensity score $p(\mathbf{L})$ is the conditional probability that $A = 1$ given $\mathbf{L}$. We can obtain this via:

- direct non-parameteric estimation when L is low dimensional and discrete, or
- by regression

$$g(E\{A|L\}) = \alpha + \beta L$$

For a given subject i, we then have that $P_i(A|L) = P(A = a_i|L = l_i)$

Inverse probability weighting

Let the weights be

$$W = \frac{1}{p(A|L)}$$

this is equal to $P(L_i)$ if $A_i = 1$ and $1 - P(L_i)$ if $A_i = 0$, Let $\widehat{W}$ be the estimates: Solving the estimating equation $\sum_i A_i \widehat{W}_i (Y_i - \theta) = 0$ gives the following solution:

$$\widehat{E}\{Y^1\} = \frac{1}{\sum A_i \widehat{W}_i} \sum_i^n \frac{A_i Y_i}{p(L_i)}$$

$$\widehat{E}\{Y^1\} - \widehat{E}\{Y^0\} = \frac{1}{\sum A_i \widehat{W}_i} \sum_i \frac{A_i Y_i}{p(L_i)} - \frac{1}{\sum (1 - A_i) \widehat{W}_i} \sum_i \frac{(1 - A_i) Y_i}{(1 - p(L_i))}$$

Assumptions:

- no interference, consistency and conditional exchangeability (given $\mathbf{L}$) hold; and
- the propensity score model has been correctly specified: i.e. in the case of the example above, that the conditional mean of A given L is correctly specified.

Dividing by n

You will see these estimators written a lot as

$$\widehat{E}\{Y^1\} = \frac{1}{n}\sum_{i}^{n}\frac{A_i Y_i}{p(L_i)}$$

This is fine because as $n \to \infty$, $\sum A_i \widehat{W_i} \to n$

Three steps for IPW

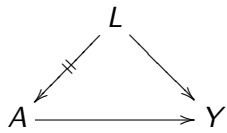
- **Step 1**: compute the probability $p(A|L)$ for each level of the exposure A and confounder L
- **Step 2**: assign a weight to each subject, equal to

$$W = \frac{1}{p(A|L)}$$

  where $p(A|L)$ is the probability of the subject's observed level of A, given the subject's observed level of L
  - for instance, suppose that $p(A = 1|L = 1) = 0.2$
  - each subject with $(A = 1, L = 1)$ is then counted as $1/0.2 = 5$ subjects
- **Step 3**: use $p(Y = 1|A = 0)$ and $p(Y = 1|A = 1)$ in the weighted sample as estimates of $p(Y^0 = 1)$ and $p(Y^1 = 1)$

Why IPW works

- IPW breaks the association between A and L

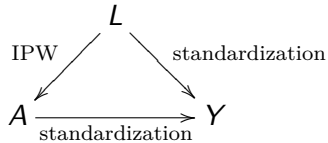


- As a consequence, L is not a confounder in the weighted sample
- If L is the only confounder in the original sample, then there is no confounder at all in the weighted sample
- Thus, in the weighted sample we have exchangeability

Outcome regression vs inverse probability weighting

- Like standardization, inverse probability weighting (IPW) is a method to estimate marginal causal effects
- Without modeling assumptions, IPW gives the same result as standardization
- IPW may give different results, and may sometimes be preferable, when using regression models

Choice of modeling assumptions



- Standardization and IPW require different models
  - standardization requires an outcome model, describing how the outcome depends on the exposure and confounders
  - IPW requires an exposure model, describing how the exposure depends on the confounders
- If we know more about how the confounders affect the exposure, then more natural to use an exposure model
- If we know more about how the confounders affect the outcome, then more natural to use an outcome model

Issues with outcome regression

- Causal inference based on traditional regression modelling requires that we model some aspect of the distribution of $Y|A, \mathbf{L}$ (usually its mean) correctly.
- Modelling $Y|A, \mathbf{L}$ might be tricky.
- If there is poor overlap, estimates will be based on extrapolations, and thus simply checking our model using the observed data will be insufficient.
- For binary/time-to-event Y, high-dimensional $\mathbf{L}$ can lead to noticeable finite sample bias.

Modeling the intervention

- Modelling $A|\mathbf{L}$ instead of $Y|A,\mathbf{L}$ can help with all these problems.
- It gives us more flexibility, since parsimony in the $A|\mathbf{L}$ model is not as desirable as it would be in the $Y|A,\mathbf{L}$ model.
- There are several ways of incorporating our model for $A|\mathbf{L}$ into the analysis: stratifying, matching, adjusting, weighting.
- These alternative methods, like traditional regression methods, are valid only if $\mathbf{L}$ is sufficient to control for all confounding.
- We also need to model $A|\mathbf{L}$ correctly, and this modelling enterprise is somewhat atypical, since including non-confounders predictive of Y is beneficial, but including non-confounders predictive of A is detrimental.

Benefits of propensity scores

- Ill-overlapping exposure groups (wrt some confounders) is a problem for causal inference no matter which method we choose. However, methods based on the propensity score flag up this problem much more strongly.
- We can sometimes get the best of both worlds by cleverly combining PS-based methods with regression adjustment. See next session on double robust methods.
- PS modelling is done without looking at the outcome: some argue that increased objectivity is an advantage.

Focusing on the ATE

There are many double robust methods for many other estimands and particularly for TMLE these other estimands can be the place where their properties are most important*. We will focus on the ATE for a

- Binary exposure A
- A continuous or binary outcome Y
- Any distribution of a set of potential confounders $\boldsymbol{L}$

Double robust (for consistency) methods

These estimators are consistent when either the propensity score OR the outcome model is correctly specified for confounding, i.e., they are statistically correctly specified and contain all confounders needed for exchangeability (i.e., a sufficient set of confounders)

Outcome regression methods

- Reminder: β in the outcome model

$$E(Y|X, \boldsymbol{L}) = \alpha + \beta A + \boldsymbol{\gamma}^T \boldsymbol{L}$$

can be given a causal interpretation as $E\left\{Y^1 - Y^0\right\}$ the ACE if:

(1) no interference, consistency and conditional exchangeability (given $\mathbf{L}$) hold; and
(2) the regression model has been correctly specified: i.e. in the case of the example above, that the conditional mean of Y given A and all the Ls is linear in each L, and that each of these 'slopes' is the same for each value of A.

Methods based on the propensity score

- The propensity score $p(\mathbf{L})$ is the conditional probability that $A = 1$ given $\mathbf{L}$ and is often modeled parametrically.

$$p(\mathbf{L}) = Pr(A = 1 | \mathbf{L}) = \text{expit}(\alpha_0 + \alpha_1 L)$$

$$\frac{1}{n} \sum_{n} \frac{A_i Y_i}{p(L_i)} - \frac{(1 - A_i) Y_i}{(1 - p(L_i))}$$

$$W(A_i, \boldsymbol{L}_i; \hat{\boldsymbol{\alpha}}) = \frac{A_i}{p(L_i)} - \frac{(1 - A_i)}{(1 - p(L_i))}$$

Assumptions:

(1) no interference, consistency and conditional exchangeability (given $\mathbf{L}$) hold; and

(2*) the propensity score model has been correctly specified: i.e. in the case of the example above, that the conditional mean of A given $\boldsymbol{L}$ is correctly specified.

Super Simple DR Method (1)

IPW Linear regression

$\hat{\beta}$ in the outcome model

$$E(Y|A, \boldsymbol{L}) = \alpha + \beta A + \boldsymbol{\gamma}^T \boldsymbol{L}$$

When we fit this model using the weighted score equations:
Then under the Assumptions in (1) and either of the Assumptions (2) or (2*), $\hat{\beta}$ is consistent for the ACE.

Super Simple DR Method (2)

Logistic Regression

$$\text{logit}\{E(Y|A=a, \boldsymbol{l}=l)\} = \gamma_0 + \beta a + \boldsymbol{\gamma}^T \boldsymbol{l},$$

When we further standardize, our estimator of the ACE is given by $\hat{E}\{\hat{E}(Y|A=1, \boldsymbol{L})\} - \hat{E}\{\hat{E}(Y|A=0, \boldsymbol{L})\}$ where

$$\hat{E}\{\hat{E}(Y|A=a, \boldsymbol{L})\} = \frac{1}{n}\sum_{i=1}^{n} \text{expit}(\hat{\gamma}_0 + \hat{\beta} a + \hat{\boldsymbol{\gamma}}^T \boldsymbol{l}_i).$$

It is consistent if either the propensity score or the outcome model is correctly specified for confounding.
Marshall Joffe discussed by Robins et al. 2007

AIPW

Regardless of how one has modeled it, let $\hat{Y}_i^a$ be your prediction for subject i under intervention $A = a$. Similarly let $\hat{p}(\boldsymbol{L}_i) = p(A = 1|\boldsymbol{L} = l)$ for a subject i regardless of how it was modelled. Then the augmented inverse probability weighting estimator is:

$$\frac{1}{n}\sum_{i=1}^{n} \frac{Y_i A_i}{\hat{p}(\boldsymbol{L}_i)} - \frac{\hat{Y}_{i1}(A_i - \hat{p}(\boldsymbol{L}_i))}{\hat{p}(\boldsymbol{L}_i)} - \frac{Y_i(1 - A_i)}{1 - \hat{p}(\boldsymbol{L}_i)} - \frac{\hat{Y}_{i0}(A_i - \hat{p}(\boldsymbol{L}_i))}{1 - \hat{p}(\boldsymbol{L}_i)}.$$

This is a consistent estimate of the ACE if the propensity or outcome models are correctly specified for confounding. But, it should now be clear that one can model $\hat{Y}_i^a$ or $\hat{p}(\boldsymbol{L}_i)$ in much more flexible ways. If you use a linear model for the outcome and logistic for the propensity score, the simple DR method will be asymptotically the same as using the AIPW with the same models.

AIPW is a one-step estimator

Note that the AIPW can be re-written as

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{A_i}{\hat{p}(\boldsymbol{L}_i)}-\frac{(1-A_i)}{(1-\hat{p}(\boldsymbol{L}_i))}\right)(Y_i-\hat{Y}_i^a)+\hat{Y}_i^1-\hat{Y}_i^0.$$

If you start with the regression estimator

$$\hat{\beta}=\sum_{n}\hat{Y}_i^1-\hat{Y}_i^0$$

and you subtract off the remainder, and estimate the true values, you get the AIPW.

Inference for the AIPW

There are three ways if we use parametric models

- nonparametric bootstrap
- Influence function-based estimator, not to be confused with EIF, of the variance that is itself DR Gabriel et al. Stat in Med (2024)
- EIF as the variance if both models are correct

Because the remainder is $o_p(n^{-1/2})$ the AIPW is are asymptotically normal and linear, meaning we can use the EIF to estimate the variance when all models are correct. This is one of the things that will be the case by construction for the TMLE.

Why not just always use the AIPW, then?

The AIPW is not a plug-in estimator, so large weights can push AIPW outside the range of the parameter. This isn't generally a big deal, but can be in smaller sample sizes and for things that are not the ACE.
TMLE is a plug-in estimator with the same asymptotic properties as the AIPW.

TMLE

Targeted maximum likelihood or targeted minimum loss estimation, with the latter being a more general version of the first, has four main steps.
For targeted maximum likelihood

1. Pick an initial estimator P^0, specifically one that makes the remainder $o_p(n^{-1/2})$
2. Define a sequence of k updates to the estimator via sub-models parameterized by ϵ, P_ϵ^k, such that the $P_\epsilon^k = P_0^k$, and the score of the submodel at $\epsilon = 0$ is the efficient influence function. We update by estimating ϵ by maximum likelihood, i.e.

$$\hat{\epsilon} = \text{argmax}_\epsilon = \sum_n \log\{P_\epsilon^k\}$$

3. Iterate until $\hat{\epsilon} \approx 0$, at K
4. the TMLE is $\Psi(P^K)$

When we allow other loss functions, the estimate of ϵ need not be obtained by maximum likelihood.

We have already seen a TMLE

The super simple DR Methods are targeted minimum loss estimators.
BUT this is not what most people think of as instead most people think of the use of the "clever covariate" method. This is a covariate that makes the submodel have a score of the EIF. In the logistic regression case, including the IPTW does the trick i.e.

$$\text{logit}(\mu_\epsilon^k)(A) = \{\text{logit}(\mu^k)(A) + \epsilon * W(A_i, \boldsymbol{L}_i; \boldsymbol{\alpha})\}$$

Running a logistic regression of Y without an intercept and including an offset variable for the initial predictions for people, we can obtain an estimate for ϵ. Then the TMLE is

$$\frac{1}{n}\sum_n \text{expit}\{\text{logit}(\hat{\mu}^0)(A) + \hat{\epsilon} * W(A_i, \boldsymbol{L}_i; \hat{\boldsymbol{\alpha}})\}$$

Why not always just use DR?

You may be taking a large risk! Getting both models wrong can be worse than getting just the outcome model wrong.
It is often better to estimate the outcome model correctly, parametrically or otherwise, if you believe you can.

Concerning trend in the applied literature

We have discussed the ACE and estimators of it for binary and continuous Y; however, regardless of the outcome, the simple combination of an outcome model and a propensity score model does not always result in a DR estimator.
An IPTW log binomial GLM, for example, does not result in a DR estimator even after standardization.

Summary ACE

- There are multiple ways to construct a DR estimator
- NOT all ways of combining a propensity score and an adjusted outcome model result in a DR estimator
- The one-step estimator and TMLE only differ in the fact that the TMLE is a plug-in estimator. This is not so important for the ACE, but can be very important for other estimands
- The AIPW and the TMLE estimators (which can take on different forms) have the same asymptotic properties if the same propensity and outcome models are used
- When you use parametric models, the inference is easy. When you use ML methods, inference can be based on the EIF if both models are correct and the ML methods are consistent at the $n^{-1/4}$ rate

Different DR estimators have different goals

- DR for missing data
- DR in survival which may be DR for misspecification of the censoring distribution
- In complex longitudinal data and/or time-varying exposures, there may be many propensity score models (NOTE that MSM are a different type of model where all models are often assumed to be correct)

References

J. Pearl. Causality: Models, Reasoning, and Inference. New York: Cambridge University Press, 2000.

Peter M. Aronow and Fredrik Sävje. The book of why: The new science of cause and effect. Journal of the American Statistical Association, 115(529):482–485, 2020.