Introduction to Causal Inference and Causal Data Science

Day 3: Addressing departures from randomisation

Bas B.L. Penning de Vries
Dept. Data Science & Biostatistics
Julius Center, UMC Utrecht

UMC Utrecht

summerschool UTRECHT

Any method may be used for confounding control:

- Regression adjustment
- (Propensity score) matching
- G-computation
- Inverse probability weighting (IPW)
- Instrumental variable analysis
- ...

Choice should be influence in part by assumptions and estimand — propensity score matching and g-methods (IPW and g-computation) target quantities typically estimated in trials

Propensity score methods

A collection of methods based on the propensity score (PS):

- PS **stratification**
- **Regression** on the PS
- PS **matching**
- Inverse probability **weighting** (IPW)

Propensity score

The propensity score ps(L) is the conditional probability of (propensity for) exposure A given a specified set L of variables:

$$\text{ps}(L) = \Pr(A = 1 \mid L)$$

- It is common practice to impose a parametric model on the propensity score — popular choice is that the *logit* of ps(L) is a linear function of the variables in L.
- *Not a must!* In modern causal inference, ps is considered a *nuisance function* — quantity used for inference but not of substantive interest + candidate for flexible *machine learning* methods (but see e.g., Alexander (2021) and Naimi et al (2021)).

Rosenbaum and Rubin (1983, Theorem 1) demonstrated **balancing property**:

conditional on ps(L), the distribution of L is the same among the treated ($A = 1$) as it is among the untreated ($A = 0$).

More importantly, they showed (Theorem 3) conditional exchangeability given L implies conditional exchangeability given the propensity score ps(L).

Key balancing property of the propensity score

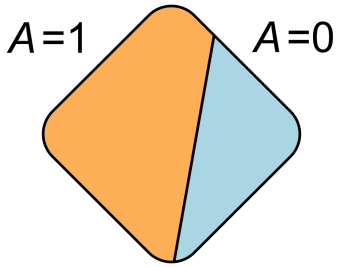
$$Y^{A=a} \perp\!\!\!\perp A \mid L \Rightarrow Y^{A=a} \perp\!\!\!\perp A \mid \text{ps}(L)$$

Powerful property: if a potentially high-dimensional L is sufficient to adjust for confounding, then so is the one-dimensional PS.

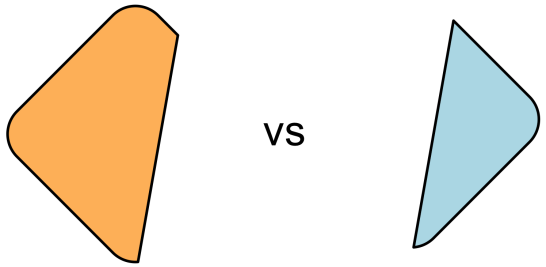
PS matching

Typical estimand is **average treatment effect among treated (ATT)**, but exact estimand depends on implementation/variation.

Population

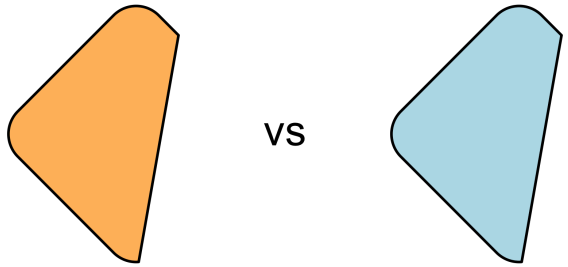


Crude comparison



$E[Y \mid A=1]$ $E[Y \mid A=0]$

ATT



$E[Y^{A=1} \mid A=1]$ $E[Y^{A=0} \mid A=1]$

- $E[Y^{A=1} \mid A = 1] = E[Y \mid A = 1]$ (consistency), but what about $E[Y^{A=0} \mid A = 1]$?
- $Y^{A=0}$ is unobserved for treated individuals — this is where (exact) matching on the propensity score comes in!

Exact propensity score matching

Exact matching on the propensity score means *replacing the unobserved outcomes* $Y^{A=0}$ of treated individuals with propensity score ps(L), with random draws Y_{match} from the observed outcomes Y of untreated individuals with the same ps(L):

$$Y_{\text{match}} \mid A = 1, \text{ps}(L) \sim Y \mid A = 0, \text{ps}(L)$$

- If we could do exact matching, it turns out that, under *conditional exchangeability, positivity and consistency*, $E[Y^{A=0} \mid A = 1] = E[Y_{\text{match}} \mid A = 1]$.
- The causal contrast $E[Y^{A=1} \mid A = 1]$ versus $E[Y^{A=0} \mid A = 1]$ is thus identified by the observed contrast $E[Y \mid A = 1]$ versus $E[Y_{\text{match}} \mid A = 1]$ — this amounts to quantifying the crude association between the treated and the matched untreated individuals.

PS matching in practice

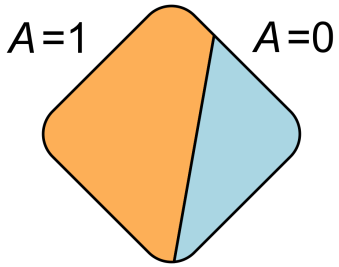
- **Problem with finite samples**: *exact matching* often (nearly always) *impossible* for all treated individuals!
- **Solution**: use approximate matching
  - Many algorithms to choose from (e.g., greedy 1:1 nearest neighbour matching with distance defined as absolute difference between logit ps(*L*); with calliper)
- **But**:
  - Allowing matching of distant subjects may result in bias (*residual confounding*).
  - Excluding subjects for whom no close-enough match can be found may result in *poor precision* (bias-variance trade-off) or *selection bias*.

- *After* propensity score estimation and matching, standard *confounding-naïve estimators* can be used to estimate the treatment-outcome effect (e.g., OLS from a simple linear regression applied to the matched treated and untreated participants).
  - *Valid for point estimation* assuming observations in the original sample are i.i.d.
  - But *naïve variance estimation is problematic*, mainly due to matching-induced outcome dependence between treated and control units through the propensity score.
- Alternative: *bootstrap*

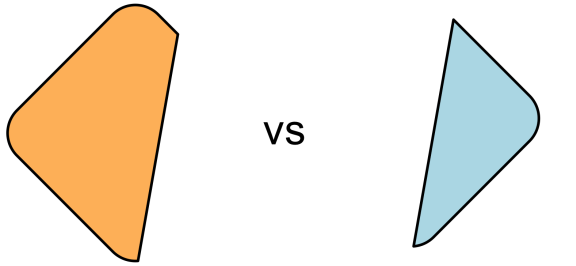
IPW for binary treatments/exposures

Typical estimand is **average treatment effect (ATE)** among everyone in the target population.

Population

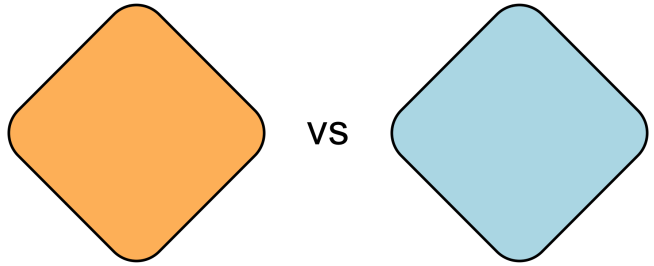


Crude comparison



$E[Y \mid A=1]$ $E[Y \mid A=0]$

ATE



$E[Y^{A=1}]$ $E[Y^{A=0}]$

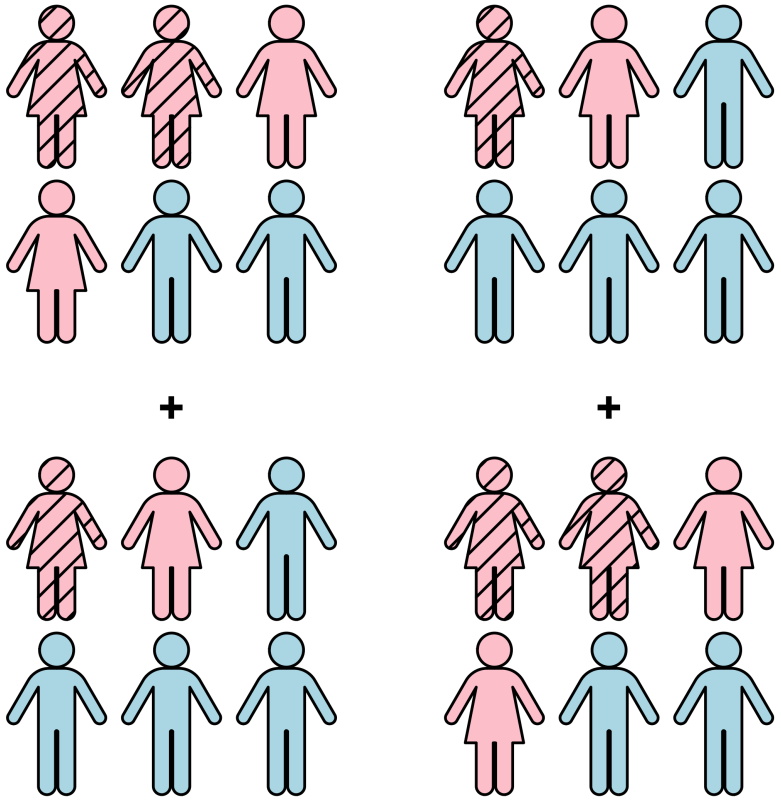
Key idea: reweight treated and untreated subpopulations so that they look (in some respects) like the entire population.

Goal: create a *pseudopopulation* (by weighting the original population) whose *treated and untreated* subgroups each have the *same distribution* of covariates L and (more importantly) *counterfactuals* $Y^{A=a}$ as the original population.

- Covariates L and counterfactual outcomes $Y^{A=a}$ are independent of treatment A in the pseudopopulation (**marginal exchangeability**).
- Causal effect = crude association, under consistency

How: it turns out that the pseudopopulation with this property can be made under *positivity* and *conditional exchangeability* given L, by ...

- weighting treated individuals with $1 / \text{ps}(L)$
- and untreated individuals with $1 / (1 - \text{ps}(L))$.



Sex	Treated	PS	Weight
Female	Yes	4/6	1.5
Female	No	4/6	3.0
Male	Yes	2/6	3.0
Male	No	2/6	1.5

Modifications

- For inference about average effect among the treated (*ATT*)
- To accommodate *more than two treatment levels*
  - Point interventions with more than two levels or interventions on *time-varying variables*
  - But with many possible treatment regimes to consider: there may be a need to make modelling assumptions regarding the distribution of the outcome and the treatment(s) in the pseudopopulation (*marginal structural modelling*)
  - 'Stabilisation' of weights recommended (to reduce their variability)
- To address selective censoring or account for *missing data*

G–computation for binary treatments

- Alternative to IPW that comes under same (non-parametric) identifiability conditions, for the same estimand (*ATE*)
- Relies on *outcome modelling* rather than treatment modelling
- Essentially "*standardisation*" of conditional quantities

Identification of average treatment effect (ATE) via g-computation

$$
\begin{aligned}
\text{ATE} &= \mathrm{E}[Y^{A=1}] - \mathrm{E}[Y^{A=0}] && \text{(causal estimand)} \\
\mathrm{E}[Y^{A=a}] &= \mathrm{E}\{\mathrm{E}[Y^{A=a} \mid L]\} && \text{(law of total expectation)} \\
&= \mathrm{E}\{\mathrm{E}[Y^{A=a} \mid L, A = a]\} && \text{(\textit{conditional exchangeability} and \textit{positivity})} \\
&= \mathrm{E}\{\mathrm{E}[Y \mid L, A = a]\} && \text{(\textit{consistency})} \\
\text{ATE} &= \mathrm{E}\{\mathrm{E}[Y \mid L, A = 1]\} - \mathrm{E}\{\mathrm{E}[Y \mid L, A = 0]\}
\end{aligned}
$$

This theoretical identification result naturally translates into the following ("g-computation") estimator:

1. **Model the conditional outcome mean**: estimate the conditional mean $E[Y \mid L, A = a]$ using an appropriate model (e.g., linear regression, logistic regression, or more flexible methods).
2. **Obtain predicted potential outcomes**: for each individual in the sample, use the model fit(s) to predict the outcome under treatment level a. This will give an estimate of $Y^{A=a}$ for everyone.
3. **Average over the empirical distribution of L**: take average as estimate of $E[Y^{A=a}]$.
4. **Contrast treatment levels**: repeat for all a and compute contrast (e.g., ATE).

The gist of g-estimation

- Originally proposed as one of three classical methods, alongside g-computation and IPW, for causal inference about the effects of time-varying treatments.
- We'll discuss the special case of time-fixed treatments ...

G-estimation for time-fixed treatments

1. Postulate model for *contrast between treatments* within levels of covariates — e.g.,

$$\mathrm{E}[Y^{A=a} - Y^{A=0} \mid L, A = a] = \beta a$$

2. Rewrite and apply *consistency*:

$$\mathrm{E}[Y^{A=0} \mid L, A = a] = \mathrm{E}[Y - \beta a \mid L, A = a]$$

3. Under *conditional exchangeability* $\mathrm{E}[Y^{A=0} \mid L, A = a] = \mathrm{E}[Y^{A=0} \mid L]$, so right-hand side should be conditionally independent of a too! Search for model parameters that are *compatible* with this:

$$\text{Search for } \beta \text{ such that } \mathrm{Cov}(Y - \beta A, A \mid L) = 0$$

International Journal of Epidemiology, 2017, 756–762
doi: 10.1093/ije/dyw323
Advance Access Publication Date: 30 December 2016
Education Corner

Education Corner

An introduction to g methods

Ashley I Naimi[1]*, Stephen R Cole[2] and Edward H Kennedy[3]

[1]Department of Epidemiology, University of Pittsburgh, [2]Department of Epidemiology, University of North Carolina at Chapel Hill and [3]Department of Statistics, Carnegie Mellon University

*Corresponding author. Department of Epidemiology University of Pittsburgh 130 DeSoto Street 503 Parran Hall Pittsburgh, PA 15261 ashley.naimi@pitt.edu

Accepted 17 October 2016

Abstract

Robins' generalized methods (g methods) provide consistent estimates of contrasts (e.g. differences, ratios) of potential outcomes under a less restrictive set of identification conditions than do standard regression methods (e.g. linear, logistic, Cox regression). Uptake of g methods by epidemiologists has been hampered by limitations in understanding both conceptual and technical details. We present a simple worked example that illustrates basic concepts, while minimizing technical complications.

Extentions of g-methods to time-varying treatment settings

- IPW, g-computation and g-estimation suitable to estimate causal effects of *time-varying treatment effects* — more on this later!

Statin-cancer example revisited (2)

Previous studies implicitly compared long-term statin users versus non-users — didn't necessarily answer questions like …

- What would be my 10-year cancer risk if I would start statin treatment now? And what if I wouldn't?
- What would be my 10-year cancer risk if I would start statin treatment now and adhered to it? And what if I wouldn't start now or in the future? (*Friday!*)

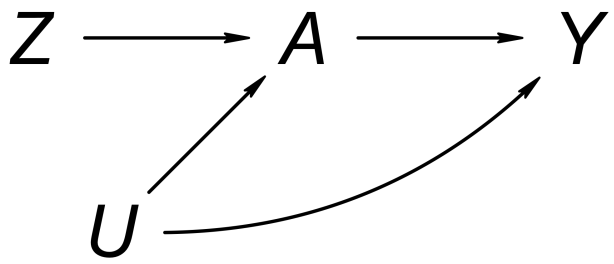
Instrumental variable analysis

- Previously discussed methods rely on the availability of a set of variables L conditional on which the *treatment groups are exchangeable*
- The key motivation for **instrumental variable analysis** is a *possible violation* of this conditional exchangeability assumption

Motivating example: encouragement trial

For inference about effect of A on Y
(or per-protocol effect), consider:

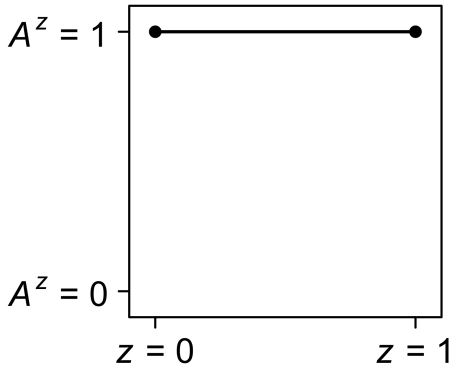
- Intention-to-treat analysis (❌ misclassification)
- As-treated analysis (❌ confounding)
- Per-protocol analysis (❌ confounding)



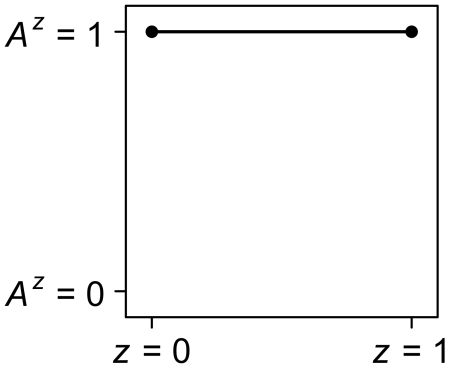
Notation

- Allocated treatment: Z
- Received treatment: A
- Received treatment had Z been z: A^z
- Outcome: Y
- Outcome had $Z = z$ and $A = a$: $Y^{z,a}$
- *Per-protocol effect*: $E[Y^{1,1}] - E[Y^{0,0}]$
- *Treatment effect*: $E[Y^{Z,1}] - E[Y^{Z,0}]$

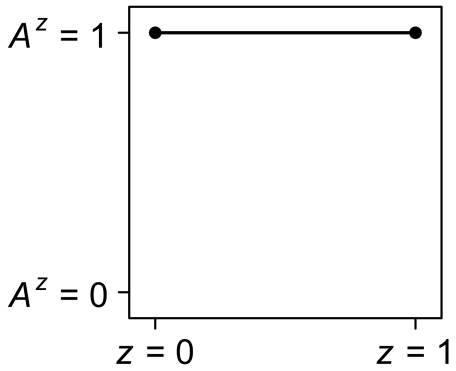
Always-takers



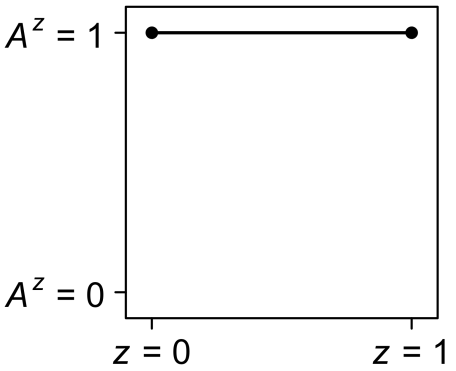
Compliers

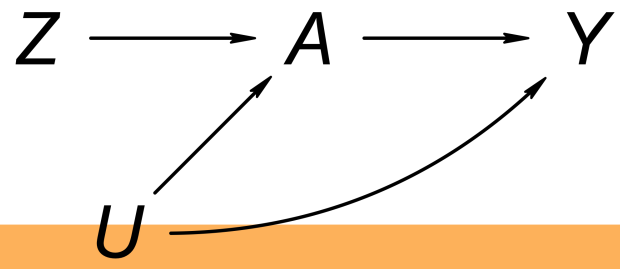


Defiers



Never-takers





Instrumental variable

Z is an **instrumental variable** (IV) if:

- $\text{Cov}(Z, A) \neq 0$ (**IV relevance**)
- $Y^{z,a}$ constant across levels of z (**IV exclusion restriction**)
- **IV exchangeability** — here: no causes shared by Z and A or Y (i.e., Y^{z,A^z} and A^z are both independent of Z, for all z)

Fourth assumption: there are **no defiers** (**monotonicity**)

Instrumental variable analysis

If, in addition, to consistency and positivity, all of the previous conditions/assumptions hold, then the **average causal difference in means among compliers** (E[$Y^{Z,1} - Y^{Z,0} \mid$ Compliers]) equals

$$\frac{E[Y \mid Z = 1] - E[Y \mid Z = 0]}{E[A \mid Z = 1] - E[A \mid Z = 0]}$$

NB: of the four IV conditions/assumptions, only the IV relevance assumption is empirically verifiable (i.e., without causal assumptions)! If compliers cannot be identified in practice, *how does IV analyis fit within TTE framework??* 🤔

Example: instrumental variable analysis

- Trial with non-adherence described by Baiocchi et al. (2014) on effect of vitamin A supplementation on child mortality
- If successful, vitamin A supplementation would not be implemented by providing pills, as in the trial, but instead by fortifying daily food items.
- 🧩 Use **IV analysis** to estimate treatment effect among the compliers.

Stratum	Assigned treatment	Received treatment	Number of children	Number of deaths
1	Control	Control	11,588	74
2	Control	Treatment	0	0
3	Treatment	Control	2,419	34
4	Treatment	Treatment	9,675	12