What is Causal Inference?

Paul Zivich, PhD

Assistant Professor
Department of Epidemiology
University of North Carolina at Chapel Hill

September 13, 2024

Acknowledgments

Open office hours: Thursdays 3-5pm Rosenau 023

pzivich@unc.edu pzivich

Thank you to Steve Cole, Jess Edwards, Bonnie Shook-Sa, Michael Hudgens, Daniel Westreich, & CIRG members

But you can only blame me for my perspective.[1]

[1] Footnotes are for references or asides. Footnotes are fair game for questions or discussion

A Series of Perspectives on Causal Inference

Observational Studies 2022 Vol 8, Issue 2

- Ian Shrier interviewed Heckman, Pearl, Robins, Rubin
- Other perspective pieces

Overview

Panoramic view rather than microscopic

- Basics
  - Causal Models
  - Parameters
  - Identification
  - Estimation
- Extensions
- Advanced Topics

Please interrupt with questions[2]

[2] Think of this as more of a guided discussion instead of a presentation

A Favorite Quote

"The subject-specific data from a longitudinal study consist of a string of numbers. These numbers represent a series of empirical measurements. Calculations are performed on these strings and causal inferences are drawn. [...] The nature of the relationship between the sentence expressing these causal conclusions and the computer calculations performed on the strings of numbers has been obscure. Since the computer algorithms are well-defined mathematical objects, it is useful to provide formal mathematical definitions for the English sentences expressing the investigator's causal inference, [...]"

James M Robins (1999 *Syntheses* 121:151-179)

What is Causal Inference to Me

A formal set of philosophical and mathematical tools and concepts that can be leveraged towards understanding systems under different sets of actions to improve decision making

Causal inference is *not* the only way to infer causal relationships

- I think 'causal inference' is an unfortunate name for the field

A Defense of 'Causal Inference'

In statistics, there is the idea of *suitable regularity conditions*. To quote Casella & Berger 2002 (pg 516), these conditions are "typically very technical, rather boring, and usually satisfied in most reasonable problems."

Some of the conditions provided are[3]

1. Observe n independent units with $X_i \sim f(x \mid \theta)$
2. Parameter is identifiable, i.e., if $\theta \neq \theta'$ then $f(x \mid \theta) \neq f(x \mid \theta')$
3. $f(x \mid \theta)$ has common support and is differentiable
4. θ is an interior point in the parameter space

[3] Ask Steve Cole about the British regularity conditions for an alternative set

A Defense of 'Causal Inference'

The field of causal inference is a reaction to this second regularity condition[4]

- No longer is identification of the parameter presumed to be 'boring' or 'usually satisfied'
- *Causal inference* as a field drags one of the background assumptions of *statistical inference* to the forefront

[4] I do not know the actual etymology. This is simply my best defense

An Algorithmic Perspective

Ontology $\rightarrow$ Parameter $\rightarrow$ Identification $\rightarrow$ Estimation

Causal Model (Ontology)

Causal Models

Some options

- Non-Parametric Structural Equation Model with Independent Errors (NPSEM-IE)[5]
- Finest Fully Randomized Causally Interpretable Structured Tree Graph (FFRCISTG)[6]
- Minimal Counterfactual Model (MCM)[7]
- (Potential Outcome) Agnostic Causal Model[8]

---

[5] Pearl 1995 *Biometrika* 82(4):669–709

[6] Robins 1986 *Mathematical Modelling* 7(9-12):1393-1512

[7] Robins & Richardson 2011 in *Causality and Psychopathology: Finding the Determinants of Disorders and their Cures*

[8] Spirtes et al. 1993 *Causation, Prediction and Search*

Implications

To progress, we need to commit to a causal model[9]

- Defines our universe of discourse
- Comes with some philosophical commitments
- Determines what parameters are well-defined objects of study

I'm going to operate under FFRCISTG

- Most discussion will also hold for NPSEM-IE
- Mostly ignore the implications

[9] Sarvet & Stensrund 2022 *Epidemiology* 33(3):372-378

Parameter

Questions to Parameters

Transform scientific question into a parameter

- Or parameter that most closely addresses the motivating scientific question

One of the most difficult steps

- Often requires revision of the question
- But a big advantage offered by causal inference
    - Can ask clearer questions (and maybe even answer them)

An Example

Does smoking cause lung cancer?

- Among who?
- When in time?
- What constitutes 'smoking'?

Potential Outcomes

To define our parameters, I will use potential variables.[10]

If Y is our outcome then

$$Y_i^a$$

is the *potential outcome*,[11] which is the value of Y for unit i if were to take action (e.g., treatment, intervention, exposure) a.[12]

- Ex: whether my clothes would be wet after arriving on campus today if I took an umbrella with me (or not)

To keep simple, will limit to a binary action A

[10] Jerzy Neyman proposed this idea in his 1923 Master's thesis
[11] FFRCISTG requires that we believe potential outcomes exist
[12] There are alternatives (e.g., do operator) but there are equivalencies

Individual Causal Effect

Potential outcome if i took action 1

$$\boxed{Y_i^1} - \boxed{Y_i^0}$$

Potential outcome if i took action 0

Here, we run into the 'fundamental problem of causal inference'.[13]

- We cannot observe the *same* unit at the *same* time and in the *same* place under two different actions
  - Even the lab scientists cannot escape this problem
  - Be suspicious of anyone who claims they can learn individual causal effects

[13] Holland 1986 *JASA* 81(396):945-960

Average Causal Effect (ACE)

Mean if took action 1

$$E[Y^1 - Y^0] = \boxed{E[Y^1]} - \boxed{E[Y^0]}$$

Mean if took action 0

where $E[\cdot]$ is the expected value function.[14]

The ACE requires a well-defined population

Unless noted otherwise, this will be the parameter of interest

[14] The subscript i's are implicit above to simplify notation

Conditional Average Causal Effect[15]

Causal effect

$$E[\, Y^1 - Y^0 \mid X = x \,]$$

among those with baseline covariates x

This parameter is like the average causal effect, but is the average causal effect among 'types' of people, where type is defined by X

- Ex: average causal effect by age

[15]When you read about people estimating the 'individual causal effect', this is usually what they are actually estimating

Stochastic Causal Effects

Mean of potential outcomes

$$\boxed{E[Y^1]}\ \boxed{\Pr^*(A=1)} + \boxed{E[Y^0]}\ \boxed{\Pr^*(A=0)}$$

Assigned probability of action

Generalization of the parts of the ACE (set $\Pr^*(A=1)=1$)

- The natural course, or $E[Y]$, can also be seen as another special case

Build contrasts between different assignments

Local Average Causal Effect

ACE on Y

$$\frac{E[Y^1 - Y^0]}{E[C^1 - C^0]}$$

ACE on compliance, C

Represents the average causal effect among those who would always comply with the specified action (an unknown subpopulation)

- Common in instrumental variable analysis
- More common in economics

Some Other Parameters

ACE among $A = 1$

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

Marginal structural models

$$E[Y^a \mid \beta]$$

Structural nested mean models

$$E[Y^a - Y^0 \mid A = a, X = x; \gamma]$$

Optimal action regimes

$$\arg\max_{r \in \mathcal{R}} E[Y^{r(a)}]$$

Risk Ratio

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

Population attributable fraction

$$\frac{E[Y] - E[Y^0]}{E[Y]}$$

Difference in CDF

$E[Y^1 \leq y] - E[Y^0 \leq y]$ for $y \in \mathcal{Y}$

Identification

Identification: Parameter Translation

Transformation of parameter into measured variables

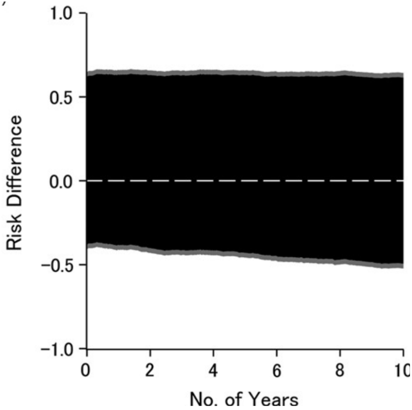
- Our parameter is expressed in terms of *potential* variables, but all we have is *observed* variables
- $Y^a \to (W, A, Y)$ for a given parameter

Variations on identification

- Partial
- Point
    - Nonparametric
    - Parametric

Partial Identification

Frechet Bounds: range of possible values of the ACE with binary Y[16] under only causal consistency (defined later)[17]



[16] Cole et al. 2019 *Am J Epidemiol* 188(4):632–636

[17] Note that these width of these bounds is always 1

Nonparametric Point Identification

Uniquely express the ACE in terms of the observed data. To do this, we will need to make assumptions about the data generating process

One *sufficient* set of assumptions is

1. Causal Consistency
2. Exchangeability
3. Positivity

Causal Consistency

Provides a connection between Y and Y^a.[18]

Set a is set to A_i for unit i

$$\text{if } \boxed{a = A_i} \text{ then } \boxed{Y_i = Y_i^a}$$

Observed outcome is *potential* outcome

[18] Cole & Frangakis 2009 *Epidemiology* 20(1):3-5

Causal Consistency

ID	A_i	Y_i	Y_i^1	Y_i^0
1	1	5	-	-
2	1	2	-	-
3	0	4	-	-
4	0	9	-	-
5	0	1	-	-

ID	A_i	Y_i	Y_i^1	Y_i^0
1	1	5	5	-
2	1	2	2	-
3	0	4	-	4
4	0	9	-	9
5	0	1	-	1

Causal Consistency Implications

Equation implies both[19]

1. Action variation irrelevance
   - Any differences in how A could be applied don't matter
   - Ex: aspirin dose on pain relief
2. No interference
   - Unit i's potential outcome does not dependent on unit j's action
   - Ex: vaccination and human-to-human transmissible diseases

[19] This special case of causal consistency is also referred to as Stable Unit Treatment Values Assumption (SUTVA)

Beyond Causal Consistency

Stochastic potential outcomes

- Distribution instead of a fixed value for i
- $E[Y^a \mid A = a] = E[Y \mid A = a]$.[20]
- Also can help to weaken action variation irrelevance

Interference[21]

- $Y_i^{\mathbf{a}} = Y_i^{a_1, a_2, \ldots, a_i, \ldots, a_n}$
- Also modifies the parameters to consider

[20] Richardson & Robins 2013 In *Second UAI Workshop on Causal Structure Learning*

[21] Hudgens & Halloran 2008 *JASA* 103(482):832-842

Exchangeability[22]

A statement about the independence of A and Y^a

These are independent...

$$Y^a \amalg A \mid W$$

...given certain covariates

which is short-hand for saying we can

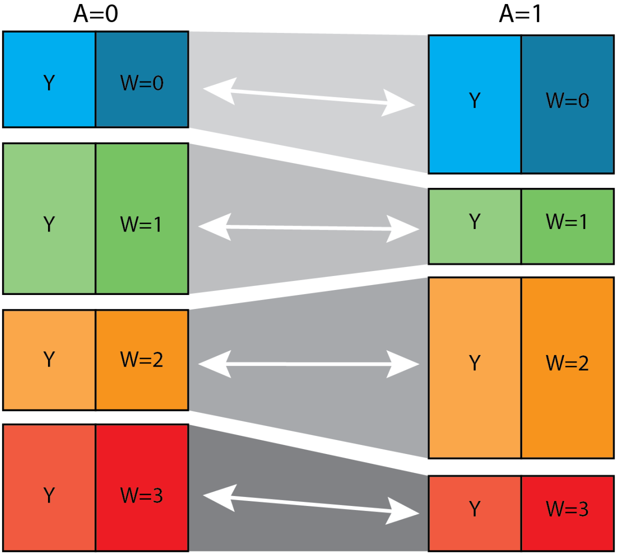
Can freely add

$$E[Y^a \mid W = w] = E[Y^a \mid \boxed{A = a}, W = w] \; \boxed{\forall (w \in \mathcal{W})}$$

all unique values of w in population

[22]You might also hear ignorability, exogeneity, no unmeasured confounding

Exchangeability Graphically



Causal Diagrams

But how do we decide what's included in W?

Causal Directed Acyclic Graphs[23]

$$
\begin{tikzcd}
W \arrow[r] \arrow[rr, bend left] & A \arrow[r] & Y
\end{tikzcd}
$$

$$
\begin{gathered}
f_W(\epsilon_W) \\
f_A(W; \epsilon_A) \\
f_Y(A, W; \epsilon_Y)
\end{gathered}
$$

[23] Greenland et al. 1999 *Epidemiology* 10(1):37-48; Lipsky & Greenland 2022 *JAMA* 327(11):1083-1084

Single World Intervention Graphs[24]

$$\begin{tikzcd} W \arrow[r] \arrow[rr, bend left] & A \mid a \arrow[r] & Y^a \end{tikzcd}$$

[24] Breskin et al. 2018 *Epidemiology* 29(3):e20-e21

Causal Diagrams for Exchangeability

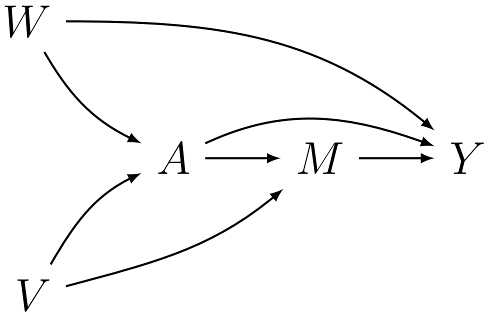
Can read independence between any two variables in a graph

- Check if any open backdoor path between two variables
  - *Closed* if there is a *collider*: $P \rightarrow Q \leftarrow R$
  - *Closed* if condition on a non-collider: $P \leftarrow \boxed{\text{Q}} \rightarrow R$
  - *Open* if condition on a collider
- Do *not* condition on 'downstream' variables

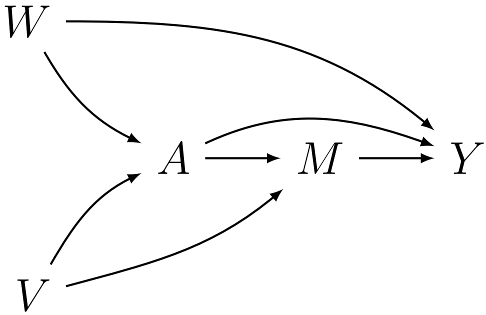
Examples

$$\begin{tikzcd} W \arrow[r] \arrow[rr, bend left] & A \arrow[r] & Y \end{tikzcd}$$

Examples



Examples

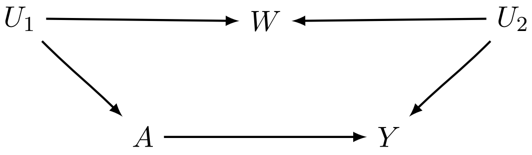


$$A \leftarrow W \rightarrow Y$$

$$A \leftarrow V \rightarrow M \rightarrow Y$$

Better Definitions[26]

Historically, epidemiologists defined a confounder as a variable associated with both the action and the outcome[25]



[25] For a better definition, see VanderWeele 2013 *Ann Stat* 41(1):196-220
[26] Causal diagrams have helped to clarify a number of concepts, such as selection bias, missing data, generalizability

Beyond Exchangeability

$$
\begin{tikzcd}
Z \arrow[dr] & & U \arrow[ll] \arrow[rr] \arrow[dl] \arrow[dr] & & W \arrow[dl] \\
& A \arrow[rr] & & Y &
\end{tikzcd}
$$

Proximal causal inference offers different identification assumptions that allow for the ACE to be identified in this context[27]

[27] Zivich et al. 2023 *Am J Epidemiol* 192(7):1224-1227, Tchetgen Tchetgen et al. 2024 *Statist Sci* 39(3):375-390

Positivity

Ensures exchangeability is mathematically well-defined

non-zero chance for $A = 1$ or $A = 0$

$$1 > \Pr(A = 1 \mid W = w) > 0 \quad \forall (w \in \mathcal{W})$$

all unique values of w in population

Exchangeability *with* Positivity[28]

In the case of binary W, Y, positivity is

$$\Pr(A = a \mid W = w) > 0 \text{ for } a \in \{0,1\} \text{ where } \Pr(W = w) > 0$$

with exchangeability needing

by def of conditional $\Pr$

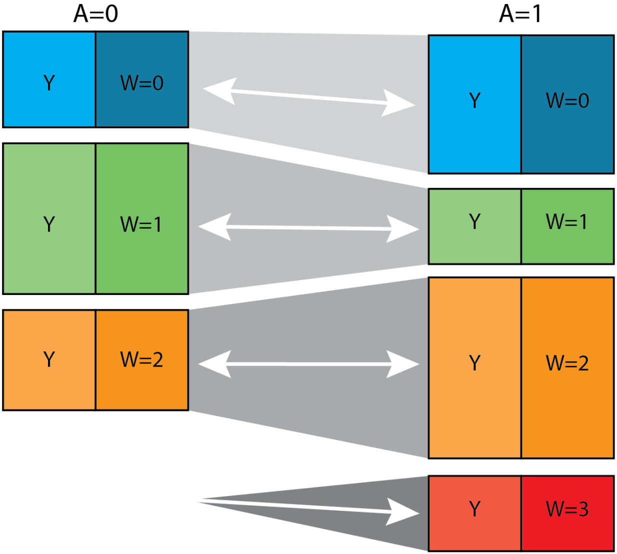
$$\Pr(Y^a = 1 \mid A = a, W = w) = \frac{\Pr(Y^a = 1, A = a, W = w)}{\Pr(A = a \mid W = w) \; \Pr(W = w)}$$

positivity ...

for W we see

[28] More of this argument is in Zivich et al. 2022 *arXiv:2207.05010*

Positivity Graphically



Beyond Positivity

Modify the population

- Restrict population to regions with positivity

Modify the parameter

- Select parameters that avoid nonpositivity
- Ex: incremental propensity score effects[29]

Leverage external information[30]

[29] Naimi et al. 2021 *Epidemiology* 32(2):202-208
[30] Greenland 2017 *EJE* 32(1):3-20,
Zivich et al. 2024 *Epidemiology* 35(1):23-31

Putting it all together...

Fundamental Theorem of Causal Inference

$$\underbrace{\Pr[Y^a = 1]}_{\text{Parameter}} = \underbrace{\sum_{w \in \mathcal{W}} \Pr[Y^a = 1 \mid W = w] \Pr(W = w)}_{\text{Law of total probability}}$$

$$= \sum_{w \in \mathcal{W}} \Pr[Y^a = 1 \mid \underbrace{A = a}_{\text{Exchangeability with Positivity}}, W = w] \Pr(W = w)$$

$$= \sum_{w \in \mathcal{W}} \Pr[\underbrace{Y}_{\text{Causal Consistency}} = 1 \mid A = a, W = w] \Pr(W = w)$$

Fundamental Theorem of Causal Inference

A more general result...

$$
\begin{aligned}
E[Y^a] =& E\left\{E[Y^a \mid W]\right\} \\
=& E\left\{E[Y^a \mid A = a, W]\right\} \\
=& E\left\{E[Y \mid A = a, W]\right\}
\end{aligned}
$$

An Equivalent Result

We can also get an equivalent expression[31]

$$
\begin{aligned}
\Pr[Y^a = 1] &= \sum_{w \in \mathcal{W}} \Pr[Y = 1 \mid A = a, W = w] \Pr(W = w) \\
&= \sum_{w \in \mathcal{W}} \frac{\Pr[Y = 1, A = a, W = w]}{\Pr(A = a \mid W = w) \Pr(W = w)} \Pr(W = w) \\
&= \sum_{w \in \mathcal{W}} \frac{\Pr[Y = 1, A = a, W = w]}{\Pr(A = a \mid W = w)} \\
&= \sum_{i=1}^{n} \frac{Y_i \; I(A_i = a)}{\Pr(A_i = a \mid W_i)}
\end{aligned}
$$

[31] This equivalence holds in the nonparametric setting

Estimation

Estimation

Once our parameter is expressed in terms of the observed data, we can leave the second regularity condition and return to statistical inference results

- But those working in causal inference have provided novel estimators
- So we will review those

Nuisance Functions

From our identification results

$$n^{-1} \sum_{i=1}^{n} E[Y_i \mid A_i = a, W_i] \qquad n^{-1} \sum_{i=1}^{n} \frac{Y_i \; I(A_i = a)}{\Pr(A_i = a \mid W_i)}$$

Requires

$$m_Y(a, W) = E[Y_i \mid A_i = a, W_i] \qquad \pi_A(W) = \Pr(A_i = a \mid W_i)$$

When these are unknown

- Estimators based on estimating one of these
- Not of direct interest, so called 'nuisance'

High-Dimensional W

In practice, W includes many variables

- Identifiable does not mean estimable[32]
- Random positivity violations
  - $\widehat{\Pr}(A = a \mid W_i = w) = 0$
- Use models to borrow information from 'nearby' observations
  - $\pi_A(W)$ replaced with $\pi_A(W; \alpha)$
  - $m_Y(a, W)$ replaced with $m_Y(a, W; \beta)$

[32] Maclaren & Nicholson 2021 *Workshop at the 38th International Conference on Machine Learning*; Aronow et al. 2021 *arXiv:2108.11342*; Robins & Ritov 1997 *Stats in Med* 16(3):285-319

Model Specification

Correct model specification

$$\Pr(A \mid W) \in \mathcal{M}_\alpha$$

Example:

$$\text{logit}\left[\pi_A(W;\alpha)\right] = \alpha_0 + \alpha_1 W + \alpha_2 W^2$$

with

$$\mathcal{M}_\alpha = \{\text{expit}(\alpha_0 + \alpha_1 W + \alpha_2 W^2) : \alpha \in \mathbb{R}^3\}$$

so $\mathcal{M}_\alpha$ covers logistic linear and quadratic relations

Propensity Scores

Donald Rubin proposed estimators based on[33]

$$\pi_A(W; \alpha)$$

Been operationalized a number of ways[34]

- Stratification
- Matching
- Regression adjustment
- Weighting

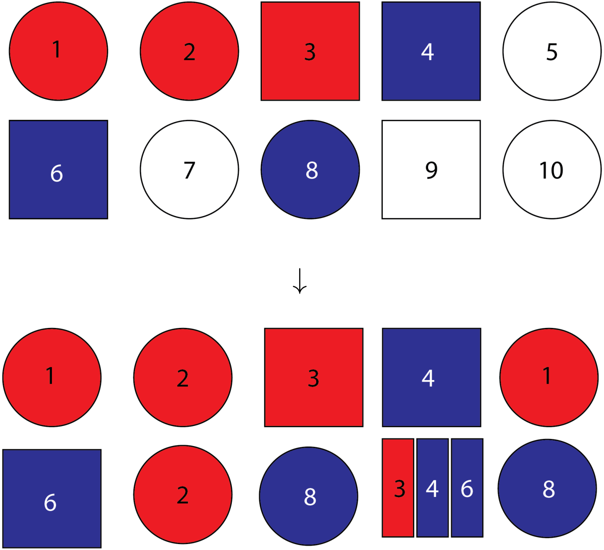
$$\hat{\mu}_{a,w} = n^{-1} \sum_{i=1}^{n} \frac{Y_i I(A_i = a)}{\pi_A(W; \hat{\alpha})}$$

---

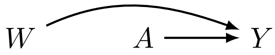
[33] Rubin 1974 *J Ed Psychol* 66:688-701

[34] Austin 2011 *Multivariate Behav Res* 46(3):399-424

Inverse Probability Weighting



IPW: Heuristically



Reweighting by $\frac{I(A_i=a)}{\Pr(A_i|W_i)}$

- Removing W, A relationship
- No confounding by W anymore

G-formula

James Robins proposed estimator based on[35]

$$m_Y(a, W; \beta)$$

$$\hat{\mu}_{a,m} = n^{-1} \sum_{i=1}^{n} m_Y(a, W; \hat{\beta})$$

[35] Robins 1986 *Mathematical Modelling* 7(9-12):1393-1512

G-formula: Heuristically

$$\begin{tikzcd} W \arrow[r] \arrow[rr, bend left] & A \arrow[r] & Y \end{tikzcd}$$

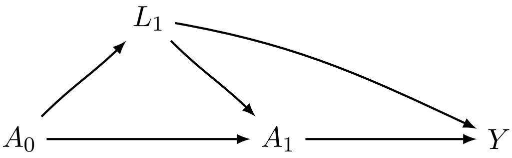
Learn $W \to Y$ and $A \to Y$ from the data

- Use that information to 'simulate' or 'impute' Y under $A := a$
- Using $m_Y(a, W)$

A Brief Aside on Time-Varying Confounding

To understand Robins's contribution, need to step back and look at time-varying actions

- Intervene on A_0 and A_1



Robins's G-Methods

G(eneral)-methods for dealing with time-varying confounding[36]

- G-formula (g-computation)
- Inverse Probability Weighting (IPW)
- G-estimation of Structural Nested Models

[36] Naimi et al. 2017 *IJE* 46(2):756-762

Augmented IPW

Rely on corresponding model to be correctly specified

AIPW cleverly combines $\pi_A(W;\alpha)$ and $m_Y(a,W;\beta)$.[37]

- Only one model needs to be correctly specified

$$n^{-1}\sum_{i=1}^{n}\left\{\overbrace{\frac{Y_i\; I(A_i=a)}{\pi_A(W;\beta)}}^{\text{IPW}} - \overbrace{m_Y(a,W;\beta)}^{\text{Outcome model}}\;\underbrace{\frac{1-\pi_A(W;\alpha)}{\pi_A(W;\alpha)}}_{\text{`Glue'}}\right\}$$

[37] A simpler implementation of AIPW is the weighted regression approach. This approach also has some performance benefits with sparse data. See Vansteelandt & Keiding 2011 *AJE* 173(7):739-742 or Shook-Sa et al. 2024 *arXiv:2404:16166*

Augmented IPW

Several important properties

- Doubly-robust
- Semiparametric efficient, unlike IPW
- Variance estimation via influence curve
- Convergence rate is a product of π_A and m_Y rates

Last one will appear again

Targeted Maximum Likelihood Estimation

Related doubly robust estimator developed by Mark van der Laan[38]

- Shares many properties with Augmented IPW
- Bounded in the parameter space

Instead combine π_A and m_Y via a targeting model

$$\text{logit}[\Pr(Y=1)] = \eta_a + \text{logit}[m_Y(a,W;\beta)]$$

estimated with weights $\frac{I(A_i=a)}{\pi_A(W_i;\alpha)}$

[38] van der Laan & Rubin 2006 *IJB* 2(1) with Schuler & Rose 2017 *AJE* 185(1):65-73 for a good introduction

Causal Effect Estimation with Machine Learning

Concern over correct model specification

- Parametric models are relatively simplistic
- Interest in more flexible alternatives to increase what $\mathcal{M}_\eta$ may cover

Promise of machine learning

- Provide data-adaptive estimation
- Bolster model specification assumption
- Does **not** deal with identification

Causal Effect Estimation with Machine Learning

Two methodological challenges[39]

1. Statistical convergence rates[40]
   - How fast estimators goes to truth by n
   - Flexibility means convergence is below $n^{-1/2}$
2. Complexity[41]
   - Certain algorithms overfit causing issues

Practical concerns, like pseudo-RNG,[42] computational time

[39] Zivich et al. 2022 *Wiley StatsRef* stat08412
[40] Daniel 2018 *Wiley StatsRef* stat08068
[41] Chernozhukov et al. 2018 *The Econometrics Journal* 21(1):C1-C68; Zivich & Breskin 2021 *Epidemiology* 32(3):393-401
[42] Zivich 2024 *Epidemiology* In-Press

Addressing Statistical Convergence[43]

Finally, if both $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\beta}}$ are estimated by ML/OLS as described earlier, then both will converge at rate $O_{\mathbb{P}}(n^{-1/2})$, meaning that the product

$$\left\{ \frac{1}{\pi(\mathbf{X}_i; \widehat{\boldsymbol{\alpha}})} - \frac{1}{\pi(\mathbf{X}_i; \boldsymbol{\alpha}^*)} \right\} \{ m(\mathbf{X}_i; \widehat{\boldsymbol{\beta}}) - m(\mathbf{X}_i; \boldsymbol{\beta}^*) \} \tag{10}$$

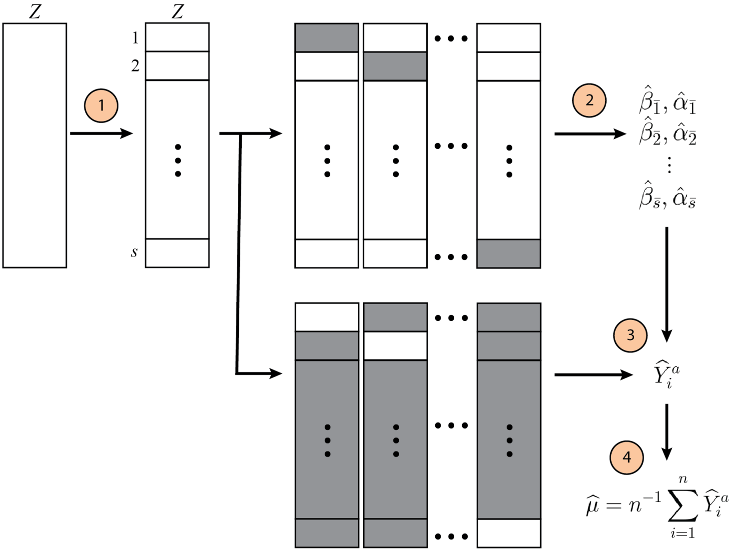
. . .

3.4 Data-Adaptive Estimation

Another consequence of the convergence result we sketched in Section 2.7.2, apart from leading to convenient inference, is that, unlike other estimators such as IPW and OR, good asymptotic properties of the resulting DR estimator can be achieved even when the convergence rates of the nuisance models is slower than the conventional parametric $\sqrt{n}$ rate. This opens the door to using data-adaptive (machine learning) estimation strategies to estimate $\pi_0(\mathbf{X})$ and $m_0(\mathbf{X})$ without incurring small sample bias, and retaining tractable inferences, as long as both estimated nuisance functionals converge to their respective truths, and that the convergence of the product shown in Expression (10) is just faster than $\sqrt{n}$[3].

[43] Screenshots from Daniel 2018 *Wiley StatsRef* stat08068

Addressing Complexity



Machine Learning Won't Always Save You

Example:

$$\Pr(A_i = 1 \mid W_i) = \begin{cases} 0.9 & \text{if } j \text{ is even} \\ 0.1 & \text{if } j \text{ is odd} \end{cases}$$

where $\mathcal{W}$ divided into $10n$ bins and j is bin ID that W_i lies in.[44]

$\therefore$ nonparametrically identified but fail to consistently estimate

- Require some smoothness of the nuisance functions

[44] Example adapted from Aronow et al. 2021 *arXiv:2108.11342*

Other Causal Approaches

Instrumental Variables[45]

Difference-in-difference and other time-series analysis[46]

Synthetic Controls[47]

Mathematical modeling[48]

Causal Discovery[49]

---

[45]Greenland 2000 *IJE* 29(4):722-729; Richardson & Tchetgen Tchetgen 2022 *AJE* 191(5):939-947

[46]Haber et al. 2021 *AJE* 190(11):2474-2486; Richardson 2023 *Epidemiology* 34(2):167-174

[47]Abadie 2021 *Journal of Economic Literature* 59(2):391–425

[48]Ackley et al. 2022 *AJE* 191(1):1-6; Murray et al. 2021 *AJE* 190(8):1652–1658

[49]Huber 2024 *arXiv:2407.08602*

Variance Estimation

Left off uncertainty estimation

- Uncertainty in parameter
- Depends on nuisance models when estimated
- Solutions
  - Bootstrap
  - Influence curve[50]
  - Sandwich variance estimator[51]

[50] Hines et al. 2022 *Am Stat* 76(3):292-304
[51] Ross et al. 2024 *IJE* 53(2):dyae030

Estimating Equations

Defined as[52]

$$E[\psi(O_i;\theta)] = 0$$

with the corresponding estimator

$$\sum_{i=1}^{n} \psi(O_i;\hat{\theta}) = 0$$

where

- ψ is a vector-valued function of dimension k
- θ is the parameter vector of dimension k

[52] Stefanski & Boos 2002 *Am Stat* 56(1):29-38

Estimating Equations

Example: IPW[53]

$$\psi_w(O_i;\theta) = \begin{bmatrix} (A_i - \pi_A(W_i;\alpha))\mathbb{W} \\ \frac{Y_i A_i}{\pi_A(W_i;\alpha)} - \mu_1 \\ \frac{Y_i(1-A_i)}{1-\pi_A(W_i;\alpha)} - \mu_0 \\ (\mu_1 - \mu_0) - \phi \end{bmatrix}$$

where $\theta = (\alpha, \mu_1, \mu_0, \phi)$

- Propensity score model
- IPW for $A := 1$
- IPW for $A := 0$
- Average causal effect

[53] Ross et al. 2024 *IJE* 53(2):dyae030 works through this and g-computation, and provides corresponding SAS/R/Python code

Inference with Estimating Equations

Sandwich variance

$$\mathbf{V}(\theta) = \mathbf{B}(\theta)^{-1}\mathbf{M}(\theta)[\mathbf{B}(\theta)^{-1}]^T$$

with the ‘bread’

$$\mathbf{B}(\theta) = E\left[\frac{\partial}{\partial\theta}\psi(O_i;\theta)\right]$$

and ‘meat’

$$\mathbf{M}(\theta) = E\left[\psi(O_i;\theta)\psi(O_i;\theta)^T\right]$$

Automated in R’s `geex` and Python’s `delicatessen`

Extensions

Causal Inference as Missing Data

Recall from causal consistency

ID	A_i	Y_i	Y_i^1	Y_i^0
1	1	5	5	-
2	1	2	2	-
3	0	4	-	4
4	0	9	-	9
5	0	1	-	1

$\therefore$ Causal inference $=$ Missing Data $+ Y^a$

Tools for Missing Data

Concepts and tools extend to any scenario we can frame as a missing data problem[54]

- Missing data
- Measurement error
- Selection bias

[54] Edwards et al. 2015 *IJE* 44(4):1452-1459

Missing Data

Causal Diagrams for Missing Data[55]

$$\begin{tikzcd} & \boxed{R_Y} \\ W \arrow[r] \arrow[ur, bend left] & Y \end{tikzcd}$$

Adapt causal estimators[56]

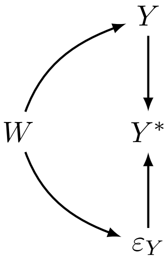
- ‘Action’ is to prevent all missingness

[55] Daniel et al. 2012 *SMMR* 21(3):243-256

[56] Vansteelandt et al. 2010 *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences* 6(1):37–48

Measurement Error

Causal Diagrams for Measurement Error[57]



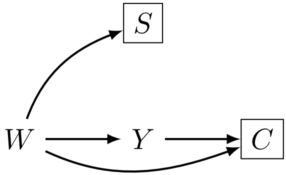
Adapt causal estimators[58]

- 'Action' is to observe Y

[57] Hernán & Cole 2009 *AJE* 170(8):959-962
[58] Ross et al. 2024 *Epidemiology* 35(2):196-207

Selection Bias

Causal Diagrams for Selection Bias[59]



Adapt causal estimators

- 'Action' is to observe Y regardless of S

[59] Lu et al. 2022 *Epidemiology* 33(5):699-706

Uses in Randomized Trials

Improved efficiency[60]

- IPW / g-computation can increase power

Per-protocol

- Break randomization so susceptible to confounding

Generalizability

- Draw inference for populations different from trial participants

[60] Morris et al. 2022 *Trials* 23(328)

Advanced Topics

Causal Mediation

Single World Intervention Graph

$$\begin{tikzcd} A \mid a \arrow[r] \arrow[rr, bend left] & M^a \mid m \arrow[r] & Y^{a,m} \end{tikzcd}$$

Mediation Parameters

Controlled direct effect

- $E[Y^{a,m}] - E[Y^{a',m}]$
- Effect of A after setting mediator fixed as m

Natural direct effect[61]

- $E[Y^{a,M(a')}] - E[Y^{a',M(a')}]$
- Effects of A with M under level where $A = a'$

Natural effects are where NPSEM-IE & FFRCISTG diverge

- Cross-world assumption

[61] Natural indirect effects instead contrast $M(a)$

Interference Parameters

Potential outcomes

$$Y_i^{a_1,\ldots,a_n} = Y_i^{a_i,a_{-i}}$$

Direct (unit-action) effect

$$E[Y_i^{a_i,a_{-i}}] - E[Y_i^{a_i',a_{-i}}]$$

Indirect (spillover) effect

$$E[Y_i^{a_i,a_{-i}}] - E[Y_i^{a_i,a'_{-i}}]$$

Total effect

$$E[Y_i^{a_i,a_{-i}}] - E[Y_i^{a_i',a'_{-i}}]$$

Overall effect

$$E[Y_i^{g(\mathbf{a})}] - E[Y_i^{g'(\mathbf{a})}]$$

Bayesian Causal Inference

Everything so far has been frequentist

Can also be Bayesian[62]

- Implications for our parameter of interest
- Use of propensity scores not motivated.[63]

[62] Li et al. 2023 *Philosophical Transactions of the Royal Society A* 381(2247):20220153

[63] Robins et al. 2015 *Biometrics* 71(2):296-299

Bayes & Propensity Scores

Likelihood of the observed data

$$\mathcal{L}(\alpha, \beta, \gamma) = \prod_{i=1}^{n} f(Y_i \mid A_i, W_i; \beta) f(A_i \mid W_i; \alpha) f(W_i; \gamma)$$

Given α, β are independent, the posterior of our parameter is entirely determined by β, γ

So a Bayesian, only evaluates

$$p(E[Y^a]) = \int_{\beta,\gamma} \left\{ \prod_{i=1}^{n} f(Y_i \mid A_i, W_i; \beta) f(W_i; \gamma) p(\beta, \gamma) \right\} \partial\beta\partial\gamma$$