Causal Inference

a summary

Contents

1 Prerequisites

1.1 Real Analysis (Abbott, 2015)

Real Numbers *triangle inequality* $|a+b| \leq |a| + |b|$
Density of Q in R: $\forall a, b \in \mathbb{R} : \exists r \in \mathbb{Q} : a < r < b$
Archimedian Property: $\forall x \in \mathbb{R}\, \exists n \in \mathbb{N} : x < n$ & $\forall y > 0\, \exists n \in \mathbb{N} : \frac{1}{n} < y$
Bounds of $A \subseteq \mathbb{R}$ upper: $\exists b \in \mathbb{R}$ s.t. $a \leq b\ \forall a \in A$ (lower: $\geq$)
least upper bound (supremum) $s \in \mathbb{R}$ s.t. s is upper bound & $\forall$ upper bounds b: $s \leq b$; greatest lower (infimum) analogous
Cardinality: $A \sim B$, if $\exists f : A \to B$, where f is 1-1 and onto
function $f : A \to B$ mapping $f(x) = ...$, domain $= A$, range $\subseteq B$
"1-1" if $a_1 \neq a_2 \Rightarrow f(a_1) \neq f(a_2)$, "onto" if $\forall b \in B\, \exists a \in A : f(a) = b$
Axiom of Completeness: every nonempty set of real numbers that is bounded above has a least upper bound; AoC, NIP, BW, CC, MCT are equivalent: if one is assumed the others follow
Nested Interval Property: if $I_n = [a_n, b_n] = \{x \in \mathbb{R} : a_n \leq x \leq b_n\}$, where $n \in \mathbb{N}$ and $I_1 \supseteq I_2 \supseteq I_3 ...$, then $\cap_{n=1}^{\infty} I_n \neq 0$

Sequences are functions with domain $\mathbb{N}$
Convergence: (a_n) converges to $a \in \mathbb{R}$ if $\forall \epsilon > 0 \exists N \in \mathbb{N}$ s.t. $n \geq N \Rightarrow |a_n - a| < \epsilon$; written as $\lim a_n = a$ or $(a_n) \to a$
Cauchy Criterion: sequence converges $\Leftrightarrow$ is Cauchy sequence
Cauchy sequence: $\forall \epsilon > 0 \exists N \in \mathbb{N}$ s.t. $m, n \geq N \Rightarrow |a_n - a_m| < \epsilon$
Boundedness: (x_n) is bounded if $\exists M > 0$ s.t. $|x_n| \leq M \forall n \in \mathbb{N}$
Algebraic Limit Theorem: if $(a_n) \to a$, $(b_n) \to b$, then $(ca_n) \to ca$, $(a_n + b_n) \to a + b$, $(a_n b_n) \to ab$, $(a_n / b_n) \to ab$ for $b \neq 0$
Order Limit Theorem: $a_n \geq 0\, \forall n \in \mathbb{N} \Rightarrow a \geq 0$ ($\leq$ analogous), $\exists c \in \mathbb{R} c \leq b_n \forall \mathbb{N} \Rightarrow c \leq b$ ($\geq$ analogous)
Monotone Convergence Theorem: bounded & monotone (*increasing* $a_n \leq a_{n+1}$ or *decreasing* $a_n \geq a_{n+1}$) sequences converge
Bolzano-Weierstrass Theorem: all bounded sequences have a convergent subsequence
subsequences of a convergent sequence converge to the same limit

Series infinite series are sums over sequences: $\sum_{n=1}^{\infty} b_n$
harmonic series $\sum_{n=1}^{\infty} \frac{1}{n}$, *geometric series* $\sum_{k=0}^{\infty} ar^k \overset{|r|<1}{=} \frac{1}{1-r}$
Convergence: to B, if $(s_m) \to B$ with partial sums $s_m = \sum_{n=1}^{m}$
Cauchy Criterion: $\sum_{k=1}^{\infty} a_k$ converges $\Leftrightarrow$
$\forall \epsilon > 0 \exists N \in \mathbb{N}$ s.t. $n > m \geq N \Rightarrow |a_{m+1} + a_{m+2} + ... + a_n| < \epsilon$; that implies if $\sum_{k=1}^{\infty} a_k$ converges then $(a_k) \to 0$
Algebraic Limit Theorem: if $\sum_{k=1}^{\infty} a_k = A$ and $\sum_{k=1}^{\infty} b_k = B$ then $\sum_{k=1}^{\infty} ca_k = cA$ and $\sum_{k=1}^{\infty} a_k + b_k = A + B$
Cauchy Condensation Test: if (b_n) is decreasing and $b_n \geq 0 \forall n \in \mathbb{N}$ then: $\sum_{n=}^{\infty} b_n$ converges $\Leftrightarrow \sum_{n=0}^{\infty} 2^n b_{2^n}$ converges
Comparison Test: if $0 \leq a_k \leq b_k \forall k \in \mathbb{N}$, then $\sum_{k=1}^{\infty} b_k$ converges $\Rightarrow \sum_{k=1}^{\infty} a_k$ too & $\sum_{k=1}^{\infty} a_k$ diverges $\Rightarrow \sum_{k=1}^{\infty} b_k$ too
Absolute Convergence Test:
$\sum_{n=1}^{\infty} |a_n|$ converges $\Rightarrow \sum_{n=1}^{\infty} a_n$ too
Alternating Series Test: if (a_n) is decreasing and converges, then $\sum_{n=1}^{\infty} (-1)^{n+1} a_n$ converges
Absolute Convergence: if $\sum_{n=1}^{\infty} |a_n|$ converges then $\sum_{n=1}^{\infty} a_n$ converges *absolutely*, if only the latter, then *conditionally*
Rearrangements: if $\sum_{k=1}^{\infty} a_k$ converges absolutely, then any rearrangement converges to the same limit
Double Series: if $\sum_{i=1}^{\infty} \sum_{j=1}^{\infty} |a_{ij}|$ converges $\Rightarrow \sum_{i=1}^{\infty} \sum_{i=1}^{\infty} a_{ij}$ $= \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} a_{ij} = \lim_{n \to \infty} s_{nn}$, where $s_{nn} = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}$

Sets *Cantor Set:* $C = \cap_{n=0}^{\infty} C_n$, with C_n removing the middle third of all intervals, e.g. $C_1 = C_0 \setminus (\frac{1}{3}, \frac{2}{3}) = [0, \frac{1}{3}] \cup [\frac{2}{3}, 1]$

Open Sets: $\forall a \in O\, \exists V_\epsilon(a) \subseteq O$, with ϵ-neighborhood of a $V_\epsilon(a) = \{x \in \mathbb{R} : |x - a| < \epsilon\}$; union of open sets is open, the finite intersecton of open sets is open
Closed Sets: contain their limit points $\Leftrightarrow$ every Cauchy sequence has a limit that lies within the set
x is *limit point* of $A \Leftrightarrow \forall \epsilon > 0\, V_\epsilon(x) \cap A$ includes other points than $x \Leftrightarrow x = \lim(a_n)$ for some $(a_n) \in A$ with $a_n \neq x \forall n \in \mathbb{N}$; all non limit point $a \in A$ are *isolated points*; the finite union of closed sets is closed, the intersection of closed sets is closed
Closure: $\bar{A} = A \cup L$, with L the set of A's limit points; the closure is the smallest closed set containing A
Complement: $A^c = \{x \in \mathbb{R} : x \notin A\}$; A closed $\Leftrightarrow$ A^c open
Compact $\Leftrightarrow$ Bounded and Closed $\Leftrightarrow$ $\exists$ Finite Subcover

- *Compactness:* K compact $\Leftrightarrow$ every sequence has a subsequence that converges in K; the intersection of a sequence of nested nonempty compact sets is not empty
- *Boundedness:* $\exists M > 0$ s.t. $|a| \leq M\, \forall a \in A$
- *Any open cover for A has a finite subcover:* An open cover is a set of open sets $\{O_\lambda : \lambda \in \Lambda\}$ whose union contains A; a finite subcover is a finite subset that still covers A

Perfection: closed and no isolated points; a nonempty perfect set is uncountable; the Cantor set is perfect
Separation: of A and B if $\bar{A} \cap B$ & $A \cap \bar{B}$
Disconnection: if $A = B \cup C$, with B, C nonempty & separated; E is connected $\Leftrightarrow$ all nonempty disjoint sets B, C s.t. $E = B \cup C$ have a convergent sequence with a limit in the other set $\Leftrightarrow$ whenever $a < c < b$ with $a, b \in E$, then $c \in E$
Baire's Theorem: $\mathbb{R}$ cannot be written as the countable union of nowhere-dense sets; E is nowhere-dense if $\bar{E}$ contains no nonempty open intervals

Functional Limits and Continuity hi
Functional Limit: let $f : A \to \mathbb{R}$ and c limit point of A
$\forall \epsilon > 0\, \exists \delta > 0$ s.t. $0 < |x - c| < \delta$ (and $x \in A$) if follows $|f(x) - L| < \epsilon$, then $\lim_{x \to c} f(x) = L$
Sequential Criterion: $\lim_{x \to c} f(x) = L \Leftrightarrow$ for all sequences $(x_n) \subseteq A$, with $x \neq c$ and $(x_n) \to c$ follows $f(x_n) \to L$
Algebraic Limit Theorem: if $\lim_{x \to c} f(x) = L$ and $\lim_{x \to c} g(x) = M$ then $\lim_{x \to c} kf(x) = kL$; $\lim_{x \to c} [f(x) + g(x)] = L + M$; $\lim_{x \to c} [f(x)g(x)] = LM$; $\lim_{x \to c} f(x)/g(x) = L/M$ if $M \neq 0$
Divergence Criterion: if (x_n) and (y_n) with $x_n \neq c \neq y_n$ and $\lim x_n = \lim y_n = c$ but $\lim f(x_n) \neq f(y_n)$ then $\nexists \lim_{x \to c} f(c)$
Continuity at c**:** $\forall \epsilon > 0\, \exists \delta > 0$ s.t. whenever $|x - c| < \delta$ (and $x \in A$) then $|f(x) - f(c)| < \epsilon$, can also be expressed as: $(x_n) \to c$ (with $x_n \in A$) $\Rightarrow f(x_n) \to f(c)$
Algebraic Continuity Theorem: if f, g continuous at c then these are too: $kf(x)$, $f(x) + g(x)$, $f(x)g(x)$, $f(x)/g(x)$ (if $g(x) \neq 0$)
Compositions: f continuous at c and g is continuous at $f(c)$ $\Rightarrow g \circ f$ is continuous at c (if $g \circ f(x)$ well-defined)
Boundedness: f is bounded on its domain $A \Leftrightarrow f(A)$ is bounded; f is bounded *on* $B \Leftrightarrow f(B)$ is bounded
Preservation of Compact Sets: K compact $\Rightarrow f(K)$ is too; if f is continuous on a compact set, f attains min/max values
Uniform Continuity: $\forall \epsilon > 0\, \exists \delta > 0$ s.t. whenever $|x - y| < \delta$ then $|f(x) - f(y)| < \epsilon$ (i.e. difference between them is bounded); f continuous on compact set $K \Rightarrow f$ uniformly continuous on K

Sequential Criterion for Nonuniformity: $\exists \epsilon_0 > 0$ and $(x_n), (y_n)$ in A s.t. $|x_n - y_n| \to 0$ but $|f(x_n) - f(y_n)| > \epsilon_0$

Intermediate Value Theorem: $f : [a, b] \to \mathbb{R}$ continuous then $f(a) < L < f(b) \Rightarrow \exists c \in (a, b)$, where $f(c) = L$

alternatively: *Preservation of Connectedness:* $f : A \to \mathbb{R}$ continuous, $E \subseteq A$ connected $\Rightarrow F(E)$ connected

Intermediate Value Property (converse of IVM): f has IVP on $[a, b]$ if $\forall x < y$ & L s.t. $f(x) < L < f(y)$ $\exists c \in (x, y)$, where $f(c) = L$ (implies continuity if f is monotone)

Discontinuity:
- *removable*: if $\lim_{x \to c} f(x)$ exists but $\neq f(c)$
- *jump*: $\lim_{x \to c^+} f(x) \neq \lim_{x \to c^-} f(x)$
- *essential*: not continuous for another reason

The Set of Discontinuous Points D_f can be written as the countable union of closed sets (=: F_σ)

2 General

Ladder Of Causation (Pearl, 2019)

1. rung:	**association**	$Pr[y\|x]$	observation	*What is?*
2. rung:	**intervention**	$Pr[y\|do(x),z]$	experiment	*What if?*
3. rung:	**counterfactuals**	$Pr[y^x\|x',y']$	retrospection	*Why?*

if a tool can answer rung i questions, it can also answer rung $j < i$

Causal Roadmap (Petersen and van der Laan, 2014)
systematic approach linking causality to statistical procedures
1. Specifying Knowledge. structural causal model (unifying counterfactual language, structural equations, & causal graphs): a set of possible data-generating processes, expresses background knowledge and its limits
2. Linking Data. specifying measured variables and sampling specifics (latter can be incorporated into the model)
3. Specifying Target. define hypothetical experiment: decide
1. variables to intervene on: one (point treatment), multiple (longitudinal, censoring/missing, (in)direct effects)
2. intervention scheme: static, dynamic, stochastic
3. counterfactual summary of interest: absolute or relative, marginal structural models, interaction, effect modification
4. population of interest: whole, subset, different population

4. Assessing Identifiability. are knowledge and data sufficient to derive estimand and if not, what else is needed?
5. Select Estimand. current best answer: knowledge-based assumptions + which minimal convenience-based asspumptions (transparency) gets as close as possible
6. Estimate. choose estimator by statistical properties, nothing causal here
7. Interpret. hierarchy: statistical, counterfactual, feasible intervention, randomized trial

Average Causal Effect $\mathrm{E}[Y^{a=1}] \neq \mathrm{E}[Y^{a=0}]$

$$\mathrm{E}[Y^a] = \sum_y y p_{Y^a}(y) \qquad \text{(discrete)}$$
$$= \int y f_{Y^a}(y) dy \qquad \text{(continuous)}$$

individual causal effect $Y_i^{a=1} \neq Y_i^{a=0}$ generally unidentifiable
null hypothesis: no average causal effect
sharp null hypothesis: no causal effect for any individual
notation A, Y: random variables (differ for individuals); a, y: particular values; counterfactual $Y^{a=1}$: Y under treatment $a = 1$
stable unit treatment value assumption (SUTVA) Y_i^a is well-defined: no interference between individuals, no multiple versions of treatment (weaker: treatment variation irrelevance)
causal effect measures typically based on means
risk difference: $\Pr[Y^{a=1}=1] - \Pr[Y^{a=0}=1]$
risk ratio: $\frac{\Pr[Y^{a=1}=1]}{\Pr[Y^{a=0}=1]}$
odds ratio: $\frac{\Pr[Y^{a=1}=1]/\Pr[Y^{a=1}=0]}{\Pr[Y^{a=0}=1]/\Pr[Y^{a=0}=0]}$
number needed to treat (NNT) to save 1 life: -1/risk difference
sources of random error: sampling variability (use consistent estimators), nondeterministic counterfactuals
association compares $E[Y|A=1]$ and $E[Y|A=0]$, **causation** compares $E[Y^{a=1}]$ and $E[Y^{a=0}]$ (whole population)

Target Trial emulating an ideal randomized experiment
explicitly formulate target trial & show how it is emulated → less vague causal question, helps spot issues
missing data problem unknown counterfactuals
randomized experiments: missing completely at random → exchangeability (= exogeneity as treatment is exogenous)
ideal randomized experiment: no censoring, double-blind, well-defined treatment, & adherence → association is causation
pragmatic trial: no placebo/blindness, realistic monitoring
PICO (population, intervention, comparator, outcome): some components of target trial
three types of causal effects:
intention-to-treat effect (effect of treatment assignment)
per-protocol effect (usually dynamic when toxicity arises)
other intervention effect (strategy changed during follow-up)
controlled direct effects: effect of A on Y not through B
natural direct effect A on Y if $B^{a=0}$ (cross-world quantity)
principal stratum effect A on Y for subset with $B^{a=0} = B^{a=1}$
crossover experiment: sequential treatment & outcome t=0, 1 individual causal effect $Y_{it}^{a_t=1} - Y_{it}^{a_t=0}$ only identifiable if: no carryover effect, effect $\perp\!\!\!\perp$ time, outcome $\perp\!\!\!\perp$ time
time zero if eligibility at multiple t (observational data): earliest, random t, all t (adjust variance with bootstrapping)
grace periods: usually treatment starts x months after first eligible, if death before: randomly assign strategy/copy into both

Identifiability Conditions hold in ideal experiments
consistency counterfactuals correspond to data $Y = Y^A$:
if $A = a$, then $Y^a = Y$ for each individual
- precise definition of Y^a via specifying a (sufficiently well-defined a maybe impossible (effect of DNA before it was discovered), relies on expert consensus)
- linkage of counterfactuals to data (a must be seen in data)

positivity $\Pr[A = a|L = l] > 0 \ \ \forall l$ with $\Pr[L = l] > 0$;
$$f_L(l) \neq 0 \Rightarrow f_{A|L}(a|l) > 0 \ \forall a, l$$
- structural violations (inference not on full population)
- random variability (smooth over with parametric models)

can sometimes be empirically verified (if all is seen in data)
exchangeability unverifiable without randomization
- *marginal:* $Y^a \perp\!\!\!\perp A \mathrel{\widehat{=}}$ randomized experiment, counterfactuals are missing completely at random (MCAR)
- *conditional:* $Y^a \perp\!\!\!\perp A|L \mathrel{\widehat{=}}$ conditionally randomized, counterfactuals are missing at random (MAR)

alternative definition: $\Pr[A = 1|Y^{a=0}, L] = \Pr[A = 1|L]$
additional conditions:
correct measurement mismeasurement of A, Y, L results in bias
correct model specification models $\overset{\text{may}}{\rightarrow}$ misspecification bias

Effect Modification A on Y varies across levels of V
null average causal effect ≠ null causal effect per subgroup
population characteristics: causal effect measure is actually "effect in a population with a particular mix of effect modifiers"
transportability: extrapolation of effect to another population (issues: effect modification, versions of treatment, interference)
effects conditional on V may be more transportable
types: additive/multiplicative scale, qualitative (effect in opposite directions)/quantitative, surrogate/causal
calculation:
- *stratify* by V then standardize/IP weight for L,
- L as *matching* factor (ensures positivity, difficult if high-dimensional L)

collapsibility: causal risk difference and ratio are weighted averages of stratum-specific risks, can not be done for odds ratio

Interaction effects of joint interventions A and E

$$\Pr\left[Y^{1,1}{=}1\right] - \Pr\left[Y^{0,1}{=}1\right] \neq \Pr\left[Y^{1,0}{=}1\right] - \Pr\left[Y^{0,0}{=}1\right]$$

A and E have equal status and could also be considered a combined treatment AE, exchangeability for both is needed
additive scale (above): ">" superadditive and "<" subadditive;
multiplicative scale: ">" super- and "<" submultiplicative
difference to effect modification: if E is randomly assigned methods coincide, but V can not be intervened on as E can
monotonicity effect is either nonnegative or nonpositive $\forall i$
sufficient component-cause framework pedagogic model
response types for binary A: helped, immune, hurt, doomed; for binary A and E: 16 types
(minimal) sufficient causes:

- (minimal) U_1 together with $A = 1$ ensure $Y = 1$
- (minimal) U_2 together with $A = 0$ ensure $Y = 1$

sufficient cause interaction: A and E appear together in a minimal sufficient cause

NPSEM *n*onp*a*ramentric *s*tructural *e*quation *m*odel

$$V_m = f_m(pa_m, \epsilon_m)$$

counterfactuals are obtained recursively, e. g. $V_3^{v_1} = V_3^{v_1, V_2^{v_1}}$
implies any variable can be intervened on
aka finest causally interpreted structural tree graph (FCISTG)
additional assumption ∩ FCISTG ⇒ causal Markov condition:

- independent errors (NPSEM-IE): all ϵ_m mutually independent
- fully randomized (FFRCISTG): $V_m^{\bar{v}_{m-1}} \perp\!\!\!\perp V_j^{\bar{v}_{j-1}}$ if $\bar{v}_{j-1}$ subvector of $\bar{v}_{m-1}$

NPSEM-IE ⇒ FFRCISTG (assume DAGs represent latter)
NPSEM-IE assume crossworld independencies → unverifiable

Causal DAG draw assumptions before conclusions

rules: arrow means direct causal effect for at least one i, absence means sharp null holds, all common causes are on the graph
neglects: direction of cause (harmful/protective), interactions
convention: time flows from left to right
causal Markov assumption: any variable (v) | its direct causes $(pa_j) \perp\!\!\!\perp$ its non-descendants $(\neg v_j) \Leftrightarrow$ Markov factorization

$$f(v) = \prod_{j=1}^{M} f(v_j | pa_j)$$

d-separation (d for directional): a pathway in a DAG is ...

- blocked if collider or conditioned on non-collider
- opened if conditioned on collider or descendent of collider

2 variables are d-separated if all connecting paths are blocked
under causal Markov: d-separation ⇒ independence
under faithfulness: independence ⇒ d-separation
faithfulness: effects don't cancel out perfectly
discovery: process of learning the causal structure; requires faithfulness, but even with it is often impossible

Noncausal DAGs (Hernán and Robins, 2023) Y^a has to be well-defined (identifiability), what about Y^l (if $L \to Y$)?

if Y^l is not well-defined, but $L \to Y$, then the graph is not causal
statistical interpretation: only $A \to Y$ is causal, the rest simply encodes conditional independencies, *but* why should a DAG corresponding to the study variables even exist then?
hidden factor: L is only a surrogate for H, with Y^h well-defined, however, L being a surrogate can introduce bias
pragmatic approach: "cause" as a primary concept which does not need explanation in terms of well-defined interventions (approach is in need of mathematical theory)

SWIGs *s*ingle *w*orld *i*ntervention *g*raphs

counterfactual graphic approach: A turns into $A|a$, the left (right) side inherits incoming (outgoing) arrows (intervention with $A = a$); all outcomes of A get a superscript a, e. g. Y^a; more than one intervention possible, dynamic strategies require additional arrows from L to a
A and Y^a are d-separated for $\mathrm{L} \to Y^a \perp\!\!\!\perp A|L$ (for FFRCISTG)

Confounding bias due to common cause of A & Y ***not in*** L

randomization prevents confounding
backdoor path: noncausal path A to Y with arrow into A
backdoor criterion: all backdoor paths are blocked by L & no descendants of A in L ⇒ conditional exchangeability
$Y^a \perp\!\!\!\perp A|L \Rightarrow L$ fulfills backdoor criterion if faithful (FFRCISTG)
confounders in observational studies: occupational factors *(healthy worker bias),* clinical decisions *(confounding by indication/channeling),* lifestyle, genetic factors *(population stratification),* social factors, environmental exposures
given a DAG, confounding is an absolute, confounder is relative
surrogate confounders in L may reduce confounding bias
negative outcome controls: if A and Y share a common cause U: measure effect for Y_0 (before treatment) and Y_1 (after), subtract (assumption of additive equi-confounding)
front door criterion using the full mediator M: $\Pr[Y^a = 1] = \sum_m \Pr[M = m|A = a] \sum_{a'} \Pr\left[Y = 1|M = m, A = a'\right] \Pr\left[A = a'\right]$

Selection Bias bias due to common effect of A & Y ***in*** L

= conditioning on collider (can't be fixed by randomization)
examples: informative censoring, nonresponse bias, healthy worker bias, volunteer bias; often M-bias ($A \leftarrow U_1 \rightarrow L \leftarrow U_2 \rightarrow Y$)
solution: target $Y^{A,C}$, AC fulfills identifiability conditions,
if competing events, interventions may not be well-defined
multiplicative survival model: $\Pr[Y{=}0|E{=}e, A{=}a] = g(e)h(a)$ → no interaction between E and A on the multiplicative scale;
if $Y = 0$ is conditionally independent, then $Y = 1$ can't be as $\Pr[Y{=}1|E{=}e, A{=}a] = 1 - g(e)h(a)$ → conditioning on a collider could be unbiased if restricted to certain levels ($Y = 0$)

Measurement Bias aka information bias

measurements X^* of variables X can be included in DAG
independent errors U if $f(U_A, U_Y) = f(U_A)f(U_Y)$
nondifferential A: if $f(U_A|Y) = f(U_A)$; Y: $f(U_Y|A) = f(U_Y)$
mismeasurement → bias, if: $A \to Y$ *or* dependent *or* differantial
reverse causation bias caused by e. g. recall bias: independent but differential A (caused by $Y \to U_A$)
misclassified treatment: assignment Z does not determine A
exclusion restriction: ensure $Z \not\to Y$, e. g. via double-blinding

- ***per-protocol effect:*** either as-treated (→ confounded) or restricted to protocol adhering individuals (→ selection bias)
- ***intention-to-treat effect*** (→ measurement bias): advantages: Z is randomized, preserves null (if exclusion restriction holds), = underpowered α-level test of the null (only if monotonicity; underpowered may be problematic if treatment safety is tested)

sometimes mismeasurement doesn't matter as the measurement itself is of interest (Hernán and Robins, 2023)

Random Variabilty quantify uncertainty due to small n

CI: e. g. Wald CI $= \hat{\theta} \pm 1.96 \times se(\hat{\theta})$, *calibrated* if it contains 95 % of estimands (>: *conservative*, <: *anticonservative*)
large sample CI: converge to 95 % vs. *small-sample:* always valid

honest: $\exists n$ where coverage $\geq 95\,\%$, *valid:* large-sample & honest
inference: either restrict inference to sample (randomization-based inference) or inference on super-population
super-population: generally a fiction, but $\rightarrow$ simple statistical properties (where does the variability of the distribution come from: assumption population is sampled from super-population)
conditionality principle: inference should be performed conditional on ancillary statistics (e.g. L-A association) as

$$\mathcal{L}(Y) = f(Y|A,L)f(A|L)f(L)$$

exactly ancillary A, L: $f(Y|A,L)$ depends on parameter of interest, but $f(A,L)$ does not share parameters with $f(Y|A,L)$
approximately ancillary: ... does not share ***all*** parameters ...
continuity principle: also condition on approximate ancillaries
curse of dimensionality: difficult to do conditionality principle

Time-Varying Treatments

compare 2 treatments
treatment history up to k: $\bar{A}_k = (A_0, A_1, ..., A_k)$
shorthand: always treated $\bar{A} = \bar{1}$, never treated $\bar{A} = (\bar{0})$
static strategy: $g = [g_0(\bar{a}_{-1}), ..., g_K(\bar{a}_{K-1})]$
dynamic strategy: $g = \left[g_0(\bar{l}_0), ..., g_K(\bar{l}_K)\right]$
stochastic strategy: non-deterministic g
optimal strategy is where $\mathrm{E}\left[Y^g\right]$ is maximized (if high is good)

Sequential Identifiability

sequential versions of
exchangability: $Y^g \perp\!\!\!\perp A_k|\bar{A}_{k-1} \;\; \forall g, k = 0, 1, ..., K$
conditional exchangeability:

$$\left(Y^g, L^g_{k+1}\right) \perp\!\!\!\perp A_k|\bar{A}_{k-1}{=}g\left(\bar{L}_k\right), \bar{L}^k \;\; \forall g, k = 0, 1, ..., K$$

positivity: $f_{\bar{A}_{k-1},\bar{L}_k}(\bar{a}_{k-1}, \bar{l}_k) \neq 0 \;\Rightarrow$

$$f_{A_k|\bar{A}_{k-1},\bar{L}_k}(a_k|\bar{a}_{k-1}, \bar{l}_k) > 0 \;\forall \left(\bar{a}_{k-1}, \bar{l}_k\right)$$

consistency:

$$Y^{\bar{a}} = Y^{\bar{a}^*} \text{ if } \bar{a} = \bar{a}^*; \qquad Y^{\bar{a}} = Y \text{ if } \bar{A} = \bar{a};$$

$$\bar{L}^{\bar{a}}_k = \bar{L}^{\bar{a}^*}_k \text{ if } \bar{a}_{k-1} = \bar{a}^*_{k-1}; \qquad \bar{L}^{\bar{a}}_k = \bar{L}_k \text{ if } \bar{A}_{k-1} = \bar{a}_{k-1}$$

generalized backdoor criterion (static strategy): all backdoors into A_k (except through future treatment) are blocked $\forall k$
static sequential exchangeability for $Y^{\bar{a}}$ (weaker version)

$$Y^{\bar{a}} \perp\!\!\!\perp A_k|\bar{A}_{k-1}, \bar{L}_k \quad \text{for } k = 0, 1, ..., K$$

sufficient to identify mean counterfactual outcome for static strategies and can be checked on SWIGS via d-separation
time-varying confounding $\mathrm{E}\left[Y^{\bar{a}}|L_0\right] \neq \mathrm{E}\left[Y|A = \bar{a}, L_0\right]$

Treatment-Confounder Feedback

$A_0 \rightarrow L_1 \rightarrow A_1$: an unmeasured U influencing L_1 and Y turns L_1 into a collider; traditional adjustment (e. g. stratification) biased: use g-methods
g-null test sequential exchangeability & sharp null true $\Rightarrow$ $Y^g{=}Y \;\forall g \;\; \Rightarrow \;\; Y \perp\!\!\!\perp A_0|L_0 \;\; \& \;\; Y \perp\!\!\!\perp A_1|A_0, L_0, L_1$; therefore: if last two independences don't hold, one assumption is violated
g-null theorem: $\mathrm{E}\left[Y^g\right] = \mathrm{E}\left[Y\right]$, if the two independences hold ($\Rightarrow$ sharp null: only if strong faithfulness (no effect cancelling))

Causal Mediation

(Hernán and Robins, 2023)
$A \longrightarrow M \longrightarrow Y$ seen as longitudinal with k_0: A and k_1: M
decompose $\mathrm{E}\left[Y^{a=1}\right] - \mathrm{E}\left[Y^{a=0}\right]$ into cross-world quantities

- pure (aka natural) direct effect (upper path)

$$\mathrm{E}\left[Y^{a=1,M^{a=0}}\right] - \mathrm{E}\left[Y^{a=0,M^{a=0}}\right]$$

- total (aka natural) indirect effect (lower path)

$$\mathrm{E}\left[Y^{a=1,M^{a=1}}\right] - \mathrm{E}\left[Y^{a=1,M^{a=0}}\right]$$

mediation formula under NPSEM-IE (requires $Y^{a=1,m} \perp\!\!\!\perp M^{a=0}$ cross-world independence)

$$\mathrm{E}\left[Y^{a=1,M^{a=0}}\right] = \sum_m \mathrm{E}\left[Y|A=1, M=m\right] \Pr\left[M=m|A=0\right]$$

interventional interpretation advocating NPSEM-IE assuming: $A \Longrightarrow N \longrightarrow M \longrightarrow Y$ / $A \Longrightarrow O \longrightarrow Y$ (thick arrows are deterministic)
no controlled direct effects: no $N \rightarrow Y$ and no $O \rightarrow M$
FFRCISTG point of view: intervention on N and O separately if decomposable (can be verified in a randomized trial), g-formula for N and O reduces to mediation formula for A

3 Models

Modeling data are a sample from the target population

estimand:	quantity of interest,	e. g. $\mathrm{E}[Y\|A=a]$
estimator:	function to use,	e. g. $\widehat{\mathrm{E}}[Y\|A=a]$
estimate:	apply function to data,	e. g. 4.1

model: a priori restriction of joint distribution/dose-response curve; *assumption:* no model misspecification (usually wrong)
non-parametric estimator: no restriction (saturated model) = *Fisher consistent estimator* (entire population data → true value)
parsimonious model: few parameters estimate many quantities
bias-variance trade-off:
wiggliness ↑ → misspecification bias ↓, CI width ↑

Variable Selection can induce bias if L includes:

(decendant of) collider:	*selection bias under the null*
noncollider effect of A:	*selection bias under the alternative*
mediator:	*overadjustment for mediators*

temporal ordering is not enough to conclude anything
bias amplification: e. g. by adjusting for an instrument Z (can also reduce bias)

Super Learning (van der Laan et al., 2007, 2011)
oracle selector: select best estimator of set of learners Z_i
discrete super learner: select algorithm with smallest cross-validated error (converges to oracle for large sample size)
super learner: improves asymptotically on discrete version
$\mathrm{logit}(Y=1|Z)=\sum_i \alpha_i Z_i$, with $0<\alpha_i<1$ and $\sum \alpha_i = 1$
weights α_i are determined inside the cross-validation; for the prediction, Z_i trained on the full data set are used
can be cross-validated itself to check for overfitting (unlikely)

Marginal Structural Models association is causation in the IP weighted pseudo-population

$$\text{associational model } \mathrm{E}[Y|A] = \text{ causal model } \mathrm{E}[Y^a]$$

step 1: estimate/model $f[A|L]$ (and $f[A]$) → get $(S)W^A$
step 2: estimate regression parameters for pseudo-population
effect modification variables V can be included (e. g. $\beta_0+\beta_1 a+\beta_2 Va+\beta_3 V$; technically not marginal anymore), $SW^A(V)=\frac{f[A|V]}{f[A|L]}$ more efficient than SW^A

3.1 Traditional Methods

Stratification calculate risk for each stratum of L
only feasible if enough data per stratum

Outcome Regression often assume no effect modification

$$\mathrm{E}\left[Y^{a,c=0}|L\right]=\beta_0+\beta_1 a+\beta_2 aL+\beta_3 L=\mathrm{E}[Y|A,C=0,L]$$

faux marginal structural model as no IP weighting/$SW^A(L)=1$
for ATE only β_1,β_2 of interest, the rest are *nuisance parameters*

Propensity Score Methods $\Pr[A=1|L]=:\pi(L)$
$\Rightarrow A \perp\!\!\!\perp L|\pi(L)$ (definition of a balancing score); can be modelled

- ***stratification:*** create strata with similar $\pi(L)$ (e. g. deciles), but the average $\pi(L)$ might still be different in some strata
- ***standardization:*** use $\pi(L)$ instead of L to standardize
- ***matching:*** find close (→ bias-variance trade-off) values of $\pi(L)$, positivity issues arise often

propensity models don't need to predict well, just ensure exchangeability (good prediction leads to positivity problems)

Instrumental Variable Estimation L unmeasured
surrogate/proxy instruments can be used
instrumental conditions:

1. ***relevance condition:*** $Z \not\perp\!\!\!\perp A$, meaning Z is associated with A (weak association (F-statistic < 10) → weak instrument)
2. ***exclusion restriction:*** Z affects Y at most through A
   (a) population level: $\mathrm{E}[Y^{z,a}]=\mathrm{E}\left[Y^{z',a}\right]$ (sometimes enough)
   (b) ***individual level:*** $Y_i^{z,a}=Y_i^{z',a}=Y_i^a$
3. ***exchangeability:*** Z and Y have no shared causes
   (a) ***marginal:*** $Y^{a,z} \perp\!\!\!\perp Z$ (typically enough)
   (b) joint: $\{Y^{z,a}; a\in[0,1], z\in[0,1]\} \perp\!\!\!\perp Z$
4. (not needed for an instrument, just the IV estimand below)
   (a) *effect homogeneity:* (i) constant effect $A\to Y\ \forall i$ (ii) constant average effect $A\to Y\ \forall A$ (iii) no additive effect modifiers (iv) additive Z-A association is constant across L
   (b) *monotonicity:* $A^{z=1}\geq A^{z=0}\ \forall i$ (more credible than 4a)

common instruments: (physician's) general preference, access to/price of A, genetic factors (Mendelian randomization)
bounds: binary outcome ATE $[-1,1]$ (width 2) $\overset{data}{\to}$ (width 1)
natural bounds need 2a,3a (width $\Pr[A{=}1|Z{=}0]+\Pr[A{=}0|Z{=}1]$)
sharp bounds require 2a,3b (narrower than natural bounds)
IV estimand ATE: intention-to-treat ÷ measure of compliance
(1,2b,3a,4a): ATE; (1,2b,3a,4b): ATE in compliers
binary Z: $\frac{\mathrm{E}[Y|Z=1]-\mathrm{E}[Y|Z=0]}{\mathrm{E}[A|Z=1]-\mathrm{E}[A|Z=0]}$, continuous Z: $\frac{Cov(Y,Z)}{Cov(A,Z)}$;
can be calculated as *two-stage-least-squares estimator:*
1. $\mathrm{E}[A|Z]$ 2. $\mathrm{E}[Y|Z]=\beta_0+\beta_1\hat{\mathrm{E}}[A|Z]$ 3. $\hat{\beta}_1$ is IV estimate
disadvantages: often leads to wide CI, small violations of conditions can lead to large biases
regression discontinuity design: if threshold in L exists which determines A perfectly + assumption of continuity in L → jump in Y at threshold is the causal effect (if no effect modification by L); a fuzzy variant also exists (Hernán and Robins, 2023)

Causal Survival Analysis time-to-event data
additional censoring due to administrative end of follow-up
competing events (often death): censoring (assume population with death abolished) or not (after death, chance of event is zero, but what is the effect of A?) → create composite event
survival quantities k is a time point, T is time of event

- *survival* at k: $\Pr[T>k]=:\Pr[D_k=0]$
- *risk* at k: $1-\Pr[T>k]=\Pr[T\leq k]=\Pr[D_k=1]$
- *hazard* at k: $\Pr[T=k|T>k-1]=\Pr[D_k=1|D_{k-1}=0]$, *hazard ratio* is paradoxical due to in-built selection bias

modeling: some options

- ***Kaplan-Meier*** aka product limit formula (nonparametric): $\Pr[D_k=0]=\prod_{m=1}^{k}\Pr[D_m=0|D_{m-1}=0]$
- parametric e. g. log hazards model:
  - use ***IP weigths*** SW^A in structural marginal model $\mathrm{logit}\Pr\left[D_{k+1}^{a,\bar{c}=\bar{0}}=0|D_k^{a,\bar{c}=\bar{0}}=0\right]=\beta_{0,k}+\beta_1 a+\beta_2 ak$
  - ***standardize*** ($\prod_k 1-$) parametric hazards model $\Pr[D_{k+1}=1|D_k=0,C_k=0,L,A]$ weighting across L

- ***structural nested cumulative failure time model (CFT):*** $\frac{\Pr[D_k^a=1|L,A]}{\Pr[D_k^{a=0}=1|L,A]} = \exp[\gamma_k(L,A;\psi)]$ (log-linear has no upper limit 1 → rare failure ↑; if ↓, use a survival model (CST)), use g-estimation like with AFT
- ***accelerated failure time model (AFT)*** with g-estimation: $T_i^a/T_i^{a=0} = \exp(-\psi_1 a - \psi_2 a L_i)$, exchangeability for C is guaranteed via artificial censoring (include only individuals who would not have been censored either way)

time-varying two options based on g-methods as examples
standardization (plug-in estimate): risk is $\Pr\left[D_{k+1}^{\bar{a},\bar{c}=\bar{0}}=1\right] =$

$$\sum_{\bar{l}_k}\sum_{j=0}^{k}\Pr\left[D_{j+1}=0|\bar{A}_j=\bar{a}_j,\bar{L}_j=\bar{l}_j,\bar{D}_j=0\right]\times$$
$$\prod_{s=0}^{j}\Big\{\Pr\left[D_s=0|\bar{A}_{s-1}=\bar{a}_{s-1},\bar{L}_{s-1}=\bar{l}_{s-1},\bar{D}_{s-1}=0\right]\times$$
$$f\left(l_s|\bar{a}_{s-1},\bar{l}_{s-1},D_s=0\right)\Big\}$$

IP weighting: fit a pooled logistic hazard model with time-varying weights $W_k^{\bar{A}} = \prod_{m=0}^{k}\frac{1}{f(A_m|\bar{A}_{m-1},\bar{L}_m)}$

3.2 G-Methods

G-Methods *g*eneralized treatment contrasts: adjust for L
- ***standardization:*** two types of g-formula
- ***IP weighting:*** (in theory) also g-formula
- ***g-estimation:*** not needed unless longitudinal

standardization and IP weighting are equivalent, ***but*** if modeled, different "no misspecification" assumptions: outcome model (standardization), treatment model (IP weighting)
big g-formula not all methods use (sequential) exchangeability
- *problem:* DAG is known, but unmeasured variables exist
- *solution:* include un- & measured variables in big g-formula → derive alternative effect identification methods using only d-separation (e. g. front door formula)

it can always be determined, if the DAG allows for identification with the big g-formula (Hernán and Robins, 2023)
censoring: measure joint effect of A and C with $\mathrm{E}\left[Y^{a,c=0}\right]$
standardization $\mathrm{E}[Y|A=a] = \int \mathrm{E}[Y|L{=}l, A{=}a, C{=}0]\, dF_L[l]$
IP weights $W^{A,C} = W^A \times W^C$ (uses n) or $SW^{A,C} = SW^A \times SW^C$ (uses $n^{c=0}$)
g-estimation only adjusts for confounding → use IP weights

time-varying censoring $\bar{C}$: monotonic type of missing data
standardization: $\int f(y|\bar{a},\bar{c}{=}\bar{0},\bar{l})\prod_{k=0}^{K} dF\left(l_k|\bar{a}_{k-1},c_{k-1}{=}0,\bar{l}_{k-1}\right)$
IP weighting:

$$SW^{\bar{C}} = \prod_{k=1}^{K+1}\frac{1\cdot\Pr\left(C_k=0|\bar{A}_{k-1},C_{k-1}=0\right)}{\Pr\left(C_k=0|\bar{A}_{k-1},C_{k-1}=0,\bar{L}_k\right)}$$

Standardization plug-in (parametric if so) g-formula

$$\mathrm{E}[Y^a] = \overbrace{\mathrm{E}\left[\mathrm{E}[Y|A{=}a,L{=}l]\right]}^{\text{conditional expectation}} = \overbrace{\int \mathrm{E}[Y|A=a,L=l]\, f_L[l]\, dl}^{\text{joint density estimator}}$$

weighted average of stratum-specific risks; unknowns can be estimated non-parametrically or modeled
no need to estimate $f_{\boldsymbol{L}}[\boldsymbol{l}]$/integrate as empirical distribution can be used: estimate outcome model → predict counterfactuals on whole dataset → average the results (→ CI by bootstrapping)
for discrete $\boldsymbol{L}$ $\mathrm{E}[Y|A=a]$ is $\sum_l \mathrm{E}[Y|A=a,L=l]\Pr[L=l]$

time-varying standardize over all possible $\bar{l}$-histories
simulates joint distribution of counterfactuals $\left(Y^{\bar{a}},\bar{L}^{\bar{a}}\right)$ for $\bar{a}$
joint density estimator (jde)

discrete: $\mathrm{E}\left[Y^{\bar{a}}\right] = \sum_{\bar{l}}\mathrm{E}\left[Y|\bar{A}=\bar{a},\bar{L}=\bar{l}\right]\prod_{k=0}^{K} f\left(l_k|\bar{a}_{k-1},\bar{l}_{k-1}\right)$

continuous: $\int f(y|\bar{a},\bar{l})\prod_{k=0}^{K} f\left(l_k|\bar{a}_{k-1},\bar{l}_{k-1}\right) dl$

for *stochastic strategies* multiply with $\prod_{k=0}^{K} f^{int}\left(a_k|\bar{a}_{k-1},\bar{l}_k\right)$

estimation (Young et al., 2011; Schomaker et al., 2019)
1. model $f\left(l_k|\bar{a}_{k-1},\bar{l}_{k-1}\right)$ and $\mathrm{E}\left[Y|\bar{A}=\bar{a},\bar{L}=\bar{l}\right]$
2. simulate data forward in time:
   at $k=0$: use empirical distribution of L_0 (observed data)
   at $k>0$: set $\bar{A}=\bar{a}$, *draw* from models estimated in 1.
3. calculate mean of $\hat{Y}_{K,i}^{\bar{a}}$ (bootstrap for CI)

iterated conditional expectation (ice)

$$\mathrm{E}\left[Y_T^{\bar{a}}\right] = \mathrm{E}\left[\mathrm{E}\left[\mathrm{E}\left[...\mathrm{E}\left[Y_T|\bar{A}_{T-1}{=}\bar{a}_{T-1},\bar{L}_T\right]...|\bar{A}_0{=}a_0,L_1\right]|L_0\right]\right]$$

estimation (Schomaker et al., 2019)
1. model inside out: $Q_T{=}\mathrm{E}\left[Y_T|\bar{A}_{T-1},\bar{L}_T\right]$ to $Q_0{=}\mathrm{E}\left[Q_1|\bar{L}_0\right]$, predict Q_t with $\bar{A}=\bar{a}$ in each step
2. calculate mean of $\hat{Q}_{0,i}^{\bar{a}}$ (bootstrap for CI)

g-null paradox even if the sharp null holds, model misspecification can lead to it being falsely rejected

Proof: for $L_0 \to A_0 \to Y_0 \to L_1 \to A_1 \to Y_1$, $\bar{a}=(a_0,a_1)$

$$\mathrm{E}\left[Y_1^{\bar{a}}\right] \overset{\text{CE}}{=} \mathrm{E}\left[\mathrm{E}\left[Y_1^{\bar{a}}|A_0{=}a_0,L_0\right]\right]$$
$$\text{(ice)} \overset{\text{CE}^*}{=} \mathrm{E}\left[\mathrm{E}\left[\mathrm{E}\left[Y_1|\bar{L},\bar{A}{=}\bar{a},Y_0\right]|A_0{=}a_0,L_0\right]\right]$$
$$\overset{\text{LIE}}{=} \mathrm{E}\left[\sum\nolimits_{l_1}\mathrm{E}\left[Y_1|A_0{=}a_0,\bar{L},Y_0\right]\Pr\left[l_1|a_0,l_0,y_0\right]\right]$$
$$\overset{\text{LIE}}{=} \sum\nolimits_{l_0}\left[\sum\nolimits_{l_1}\mathrm{E}\left[Y_1|A_0{=}a_0,\bar{L},Y_0\right]\Pr\left[l_1|a_0,l_0,y_0\right]\right]\Pr[l_0]$$
$$\text{(jde)} \overset{\text{sum}}{=} \sum\nolimits_{\bar{l}}\mathrm{E}\left[Y_1|A_0{=}a_0,\bar{L},Y_0\right]\Pr\left[l_1|a_0,l_0\right]\Pr[l_0]$$

CE: conditional expectation; *: exchangeability;
LIE: law of iterated expectation

IP Weighting *i*nverse *p*robability of treatment (g-formula)

$$\mathrm{E}[Y^a] = \mathrm{E}\left[\frac{I(A=a)Y}{f[A|L]}\right]; W^A = \frac{1}{f[A|L]}; SW^A = \frac{f(A)}{f[A|L]}$$

unknowns can be estimated non-parametrically or modeled
pseudo-population: everyone is treated & untreated ($L \not\to A$)
FRCISTG *(fully randomized causally interpreted structured graph)*: probability tree for $L \to A \to Y$, can be used to calculate/visualize simulation of values for A
for discrete $\boldsymbol{A}$, $\boldsymbol{L}$: $f[a|l] = \Pr[A=a, L=l]$
estimators: Horvitz-Thompson; Hajek (modified version)
stabilized weights $\boldsymbol{SW^A}$ should have an average of 1 (check!)
→ pseudo-population same size → CI width ↓

time-varying

$$W^{\bar{A}} = \prod_{k=0}^{K}\frac{1}{f\left(A_k|\bar{A}_{k-1},\bar{L}_k\right)};\quad SW^{\bar{A}} = \prod_{k=0}^{K}\frac{f\left(A_k|\bar{A}_{k-1}\right)}{f\left(A_k|\bar{A}_{k-1},\bar{L}_k\right)}$$

G-Estimation (additive) structural nested models

$$\text{logit}\Pr\left[A=1|H(\psi^\dagger),L\right]=\alpha_0+\alpha_1 H(\psi^\dagger)+\alpha_2 L$$

$$H(\psi^\dagger)=Y-\psi_\dagger A$$

find $\psi^\dagger$ which renders $\alpha_1=0$; 95 %-CI: all $\psi^\dagger$ for which $p>0.05$
closed-form solution for linear models
derivation: $H(\psi^\dagger)=Y^{a=0}$

$$\text{logit}\Pr\left[A=1|Y^{a=0},L\right]=\alpha_0+\alpha_1 Y^{a=0}+\alpha_2 L$$

$Y^{a=0}$ unknown, but because of exchangeability α_1 should be zero

$$Y^{a=0}=Y^a-\psi_1 a$$

equivalent to $Y^{a=0}=Y^{a=1}-\psi_1$, but using no counterfactuals
structural nested mean model

$$\text{additive:}\quad \mathrm{E}\left[Y^a-Y^{a=0}|A=a,L\right]=\beta_1 a\,(+\beta_2 aL)$$

$$\text{multiplicative:}\quad \log\left(\frac{\mathrm{E}\left[Y^a|A=a,L\right]}{\mathrm{E}\left[Y^{a=0}|A=a,L\right]}\right)=\beta_1 a\,(+\beta_2 aL)$$

multiplicative is preferred if Y always positive, but does not extend to longitudinal case
semi-parametric: agnostic about β_0 and effect of L → robust ↑
no time-varying: no nesting; model equals marginal structural models with missing β_0,β_3 (unspecified "no treatment")

sensitivity analysis: unmeasured confounding ($\alpha_1\neq 0$) can be examined: do procedure for different values of α_1 → plot α_1 vs. $\psi^\dagger$ → how sensitive is estimate to unmeasured confounding?
effect modification: add V in both g-estimation equations
doubly robust estimators exist

time-varying nested equations: for each time k
strutural nested mean models separate effect of each a_k

$$\mathrm{E}\left[Y^{\bar a_{k-1},a_k,\underline 0_{k+1}}-Y^{\bar a_{k-1},\underline 0_{k+1}}|\bar L^{\bar a_{k-1}}=\bar l_k,\bar A_{k-1}=\bar a_{k-1}\right]=$$

$$a_k\gamma_k\left(\bar a_{k-1},\bar l_k,\beta\right)$$

calculations

$$H_k\left(\psi^\dagger\right)=Y-\sum_{j=k}^{K}A_j\gamma_j\left(\bar A_{j-1},\bar L_j,\psi^\dagger\right)$$

function γ_j can be, e. g. constant (ψ_1), time-varying only ($\psi_1+\psi_2 k$), or dependent on treatment/covariate history

$$\text{logit}\Pr\left[A_k=1|H_k\left(\psi^\dagger\right),\bar L_k,\bar A_{k-1}\right]=$$

$$\alpha_0+\alpha_1 H_k\left(\psi^\dagger\right)+\alpha_2 w_k\left(\bar L_k,\bar A_{k-1}\right)$$

find α_1 that is closest to zero
a closed form estimator exists for the linear case

3.3 Doubly Robust Methods

Double-Robustness (Hernán and Robins, 2023)
g-formula: *either* treatment model $f(L)$ *or* outcome model $b(L)$ ***or*** appropriately combine both: "two chances to get it right"
all doubly robust estimators

- involve a correction of outcome $\hat b(L)$ using the treatment $\hat f(L)$
- have a bias depending on a product of the errors $\frac{1}{\pi(l)}-\frac{1}{\hat\pi(l)}$ and $b(l)-\hat b(l)$ known as second order bias

time-varying: multiple robustness for $k=0,1,...K$
$K+2$ robustness: consistent, if $\hat f_0$ to $\hat f_l$ and $\hat b_{l+1}$ to $\hat b_K$ are
2^{K+1} robustness: consistent, if for each k, either $\hat f_k$ or $\hat b_k$ are

Machine Learning L is high-dimensional
one could use lasso or ML for IP weighting/standardization
but: ML does not guarantee elimination of confounding and has largely unknown statistical properties: how to get CI?
sample splitting: train estimators on training sample T_r, use resulting estimators for doubly robust method on estimation sample (CIs on estimation sample are valid, but n halved)
cross-fitting: do again the other way round, average the two estimates, get CI via bootstrapping [*alternatively:* split into M samples, use one sample for estimation and $M-1$ for training → improved finite sample behavior (Hernán and Robins, 2023)]
asymptotic behavior for valid (Wald) CI we need:

- a bias much smaller than $c\cdot 1/\sqrt{n}$, which is how the *se* typically scales (use doubly robust methods for small bias)
- asymptotic normality (for Wald CI)
- for a doubly robust estimator ψ_{dr}, we need sample splitting, otherwise $\hat b(l)$ and $\hat f(l)$ are correlated with ψ_{dr}

if $\hat b(l)$ and $\hat f(l)$ are consistent and $\mathrm{E}[\hat\psi-\psi|T_r]/se(\hat\psi)$ converges to $0\to\hat\psi$ with sample splitting is asymptotically normal and unbiased → CI is calibrated (Hernán and Robins, 2023)
problems: unclear choice of algorithm, is bias small enough?

Advantages (van der Laan et al., 2011)
consistent *if either* $\bar Q_0$ *or* g_n *are consistent (doubly robust)*:

$$\forall\epsilon>0,P\in\mathcal{M}:\Pr_P\left[|\hat\theta_n-\theta(P)|>\epsilon\right]\to 0\text{ as }n\to\infty$$

collaboratively doubly robust: g_n only needs predictors of Y, as it does not try to fit g_0 well, but improve the fit of $\bar Q_n^*$
asymptotic unbiasedness *if either* $\bar Q_0$ *or* g_0 *are consistent*, super learning makes $\bar Q_0$ and g_n max. asymptotically unbiased
asymptotic efficiency *if both* $\bar Q_0$ *and* g_n *are consistent*: achieves Cramer-Rao bound of minimum possible asymptotic variance (requires asymptotic unbiasedness)
asymptotic linearity *if either* $\bar Q_0$ *or* g_n *are consistent*: means estimator behaves like empirical mean

- bias converges to zero at rate smaller than $1/\sqrt{n}$
- for large n estimator is approximately normally distributed

Influence Curve how robust is an estimator?

$$IC_{T,P_n}(O)=\lim_{\epsilon\to 0}\frac{T\left[(1-\epsilon)P_n+\epsilon\delta_O\right]-T(P_n)}{\epsilon}$$

for estimator T and distribution P_n with $0<\epsilon<1$
can also be rewritten as a ***directional derivative*** at P_n

$$IC_{T,P_n}=\frac{d}{d\epsilon}T\left[(1-\epsilon)P_n+\epsilon\delta_O\right]=\frac{d}{dP_n}T\left(\delta_O-P_n\right)$$

in direction (δ_O-P_n), where P_n empirical probability measure that puts mass $1/n$ on O_i (Hampel, 1974)
special cases (van der Laan et al., 2011)

- $\overline{IC}(P_0)=0$ and $\mathrm{Var}(IC(P_0))$ asymptotic variance of the standard estimator $\sqrt{n}(\psi_n-\psi_0)$, $\to Var(\hat\Psi(P_n))=\frac{Var_{IC}}{n}$
- efficient IC: an estimator is asymptotically efficient ⇔ its influence curve is the efficient influence curve $IC(O)=D^*(O)$

Delta Method (Zepeda-Tello et al., 2022) estimand is a function of θ, i.e. $\psi:=\phi(\theta)$, $Var(\hat\theta)$ known, but what is $Var(\hat\psi)$?
Taylor's approximation requirements:

- univariate ϕ: differentiable at θ
- multivariate ϕ: $\exists\ \partial_v\phi(\theta)$ (directional derivative)
- functional ϕ (function of functions): $\exists\ \partial_v\phi(\theta)$ & coincides with one-sided directional (Hadamard) derivatives ($\stackrel{*}{=}\nabla\phi(\theta)^T v$)

first order Taylor (rearranged†): $\phi(\hat\theta_n)\ \bar{\approx}\ \phi(\theta)\ \bar{\bar{+}}\ \partial_{v:=\hat\theta-\theta}\phi(\theta)$
classical delta method: if $\{r_n\}_{n=1}^{\infty}$ with $\lim_{n\to\infty}r_n=\infty$,

where $r_n(\hat{\theta}_n - \theta)$ converges to $Z \sim \mathrm{N}(0,1)$ (e. g. $r_n = \sqrt{n/\sigma^2}$), then

$$r_n\left(\phi(\hat{\theta}_n) - \phi(\theta)\right) \overset{\dagger *}{\approx} \nabla\phi(\theta)^T r_n(\hat{\theta}_n - \theta) \overset{d}{\to} \nabla\phi(\theta)^T Z$$

$$\Rightarrow \mathrm{Var}\left[\phi(\hat{\theta}_n) - \phi(\theta)\right] = \mathrm{Var}\left[\phi(\hat{\theta}_n)\right] \approx \tfrac{1}{r_n^2}\mathrm{Var}\left[\nabla\phi(\theta)^T Z\right]$$

functional delta: $r_n(\hat{\theta}_n - \theta) \overset{d}{\to} Z \Rightarrow r_n\left(\phi(\hat{\theta}_n) - \phi(\theta)\right) \overset{d}{\to} \partial_Z \phi(\theta)$

influence function: $\psi = \phi(\mathbb{P}_X)$ is a functional estimations rate of change for $\mathbb{P}_X$ to Q, where $Q = \mathbb{1}_{\{Y\}}$

$$\mathrm{IF}_{\phi,\mathbb{P}_X}(Y) := \partial_{Q-\mathbb{P}_X}\phi(\mathbb{P}_X) = \lim_{h \downarrow 0} \frac{\phi\left((1-h)\mathbb{P}_X + hQ\right) - \phi(\mathbb{P}_X)}{h},$$

interpretation: rate of change if distribution deviates from $\mathbb{P}_X$ to $Q =$ one observation Y, assigns probability 1 to X taking value Y

use delta: $\phi(\hat{\mathbb{P}}_X) \approx \phi(\mathbb{P}_X) + \mathrm{IF}_{\phi,\mathbb{P}_X}(Y)$, if $(\hat{\theta}_n - \theta) \overset{n\to\infty}{\sim} \mathrm{N}(.,.)$

$$\hat{\psi}_n - \psi = \phi(\hat{\theta}_n) - \phi(\theta) \overset{\text{approx}}{\sim} \mathrm{N}\left(0, \mathrm{Var}[\mathrm{IF}_{\phi,\mathbb{P}_X}(Y)]\right),$$

where $\widehat{\mathrm{Var}}[\mathrm{IF}_{\phi,\mathbb{P}_X}(Y)] = \frac{1}{n}\sum_{i=1}^n \left(\mathrm{IF}_{\phi,\mathbb{P}_X}(X_i)\right)^2$, which is the classical S^2 estimator since the mean is known ($= 0$)

using the delta method (general case)

1. determine asymptotic distribution of $v := r_n(\hat{\theta}_n - \theta)$
2. define ϕ and compute Hadamard derivative
3. multiply asymptotic distribution with Hadamard derivative, then estimate the variance

Simple Plug-In Estimator proto-TMLE

1. fit outcome regression with variable $R = \begin{cases} +W^A & \text{if } A{=}1 \\ -W^A & \text{if } A{=}0 \end{cases}$
2. standardize by averaging

time-varying $K+2$ robust estimator (related to TMLE)

1. estimate $\hat{f}\left(A_m | \bar{A}_{m-1}, \bar{L}_m\right)$ (e. g. logistic model), use it to calculate at each time m: $\widehat{W}^{\bar{A}_m} = \prod_{k=0}^m \frac{1}{\hat{f}(A_k|\bar{A}_{k-1},\bar{L}_k)}$ and modified IP weights at m: $\widehat{W}^{\bar{A}_{m-1},a_m} = \frac{\widehat{W}^{\bar{A}_{m-1}}}{\hat{f}(a_m|\bar{A}_{m-1},\bar{L}_m)}$
2. with $\widehat{T}_{K+1} := Y$, recursively for $m = K, K-1, ..., 0$:
   (a) fit outcome regression on $\widehat{T}_{m+1}$ with variable $\widehat{W}^{\bar{A}_m}$
   (b) calculate $\widehat{T}_m$ using the outcome model with $\widehat{W}^{\bar{A}_{m-1},a_m}$
3. calculate standardized mean outcome $\widehat{\mathrm{E}}[Y^{\bar{a}}] = \mathrm{E}\left[\widehat{T}_0\right]$

Augmented IPTW (Hernán and Robins, 2023)

$$\hat{\mathrm{E}}[Y^a] = \frac{1}{n}\sum_{i=1}^n \left[\frac{\mathbb{1}(A=a)Y}{\hat{f}(A|L)} - \left(\frac{\mathbb{1}(A=a)}{\hat{f}(A|L)} - 1\right)\hat{b}(a,L)\right]$$

disadvantages: ignores global constraints → often unstable if sparsity, sometimes not well-defined (van der Laan et al., 2011)

Relationship between AIPTW and TMLE for causal effect:

$$\hat{\psi}_{1,AIPTW} - \hat{\psi}_{0,AIPTW} = P_n\left[\hat{b}(1,L)\right] - P_n\left[\hat{b}(0,L)\right]$$

$$-P_n\left[\frac{\{\mathbb{1}(A{=}1) - \mathbb{1}(A{=}0)\}\left(Y - \hat{b}(A,L)\right)}{\hat{f}(A|L)}\right]^{\dagger}$$

using the IRLS estimate for $b(A,L;\beta,\theta) = \phi\left[m(A,L;\beta) + \theta\left\{\overbrace{\frac{\mathbb{1}(A=1)-\mathbb{1}(A=0)}{\hat{f}(A|L)}}^{\text{clever covariate}}\right\}\right]$ with canonical link ϕ sets the last part† to zero (as the score equation for θ)

TMLE (van der Laan and Rubin, 2006; van der Laan et al., 2011)

*t*argeted *m*aximum *l*ikelihood *e*stimation: an ML-based substitution estimator of the g-formula

$$O = (W, A, Y) \sim P_0; \quad \mathcal{L}(O) = \overbrace{\Pr(Y|A,W)}^{Q}\overbrace{\Pr(A|W)}^{g}\overbrace{\Pr(W)}^{Q_W}$$

target $\Psi(P_0) = \Psi(\bar{Q}_0, Q_{W,0}) = \psi_0$, *ATE:* $\bar{Q}_0 = \mathrm{E}_0(Y|A,W)$

first step: outcome model $\bar{Q}_n^0(A,W)$ estimating $\bar{Q}_0$ (part of P_0)

- super learning is often used here, but leads to a biased estimate
- not all of P_0 is estimated, just relevant portion $\bar{Q}_0$ → efficiency

second step: update $\bar{Q}_n^0(A,W)$ to $\bar{Q}_n^1(A,W)$ using treatment model g_n estimating $g_0 = P_0(A|W)$, *e. g. for binary A:*

1. model g_n, super learning is a popular choice here, too
2. calculate n clever covariates: $H_n^*(A,W) = \begin{cases} \frac{1}{g_n(1|W)} & \text{if } A_i{=}1 \\ \frac{-1}{g_n(0|W)} & \text{if } A_i{=}0 \end{cases}$
3. update $\bar{Q}_n^0$, by estimating ϵ_n with offset logistic regression: $\mathrm{logit}\,\bar{Q}_n^1(A,W) = \mathrm{logit}\,\bar{Q}_n^0(A,W) + \epsilon_n H_n^*(A,W)$ (converges after first update), then calculate counterfactuals

- goal: bias reduction, get optimal bias-variance trade-off
- removes all asymptotic bias, if consistent estimator is used here

third step: use empirical distribution for $Q_{W,0}$ in a substitution estimator, *e. g.*: $\psi_n^{TMLE} = \frac{1}{n}\sum_{i=1}^n \left[\bar{Q}_n^1(1,W_i) - \bar{Q}_n^1(0,W_i)\right]$

advantages: loss-based (does not only solve efficient influence curve estimating equation, but also uses a loss and working model preserving global constraints), well-defined (as a loss-based learner), substition estimator (respects global constraints → more robust to outliers and sparsity) → good finite sample performance

closed form inference based on the influence curve, *e. g. :*

$$IC_n^*(O_i) = \overbrace{\left[\frac{\mathbb{1}(A_i=1)}{g_n(1,W_i)} - \frac{\mathbb{1}(A_i=0)}{g_n(0,W_i)}\right]\left[Y - \bar{Q}_n^1(A_i,W_i)\right]}^{a}$$

$$+ \overbrace{\bar{Q}_n^1(1,W_i) - \bar{Q}_n^1(0,W_i) - \psi_{TMLE,n}}^{b}$$

TMLE sets the mean of the IC, $\overline{IC}_n$, to zero (b has already mean zero, see third step, MLE sets the sum of a to zero, if $H_n^*(A,W)$ is chosen correctly → the first part of a is the clever covariate)

sample variance is then: $S^2(IC_n) = \frac{1}{n}\sum_{i=1}^n \left(IC_n(o_i) - \overline{IC}_n\right)^2$

standard error of estimator: $\sigma_n = \sqrt{\frac{S^2(IC_n)}{n}}$

95% CI: $\psi_{TMLE,n} \pm z_{0.975}\frac{\sigma_n}{\sqrt{n}}$; p-value: $2\left[1 - \Phi\left(\left|\frac{\psi_{TMLE,n}}{\sigma_n/\sqrt{n}}\right|\right)\right]$

time-varying LTMLE (Schomaker et al., 2019; van der Laan and Gruber, 2012) *l*ongitudinal TMLE: based on ice g-formula for $t = T, ..., 1$:

1. model $\hat{\mathrm{E}}(Y_t|\bar{A}_{t-1},\bar{L}_t)$ (for individuals observed at $t{-}1$)
2. plug in $\bar{A}_{t-1} = \bar{d}_{t-1}$; use regression from step 1 to predict outcome at time t, i. e. $\bar{Y}_t^{\bar{d}_t}$
3. update estimate with $\bar{Y}_{t,\text{new}}^{\bar{d}_t} = \mathrm{offset}(\bar{Y}_t^{\bar{d}_t}) + \epsilon\hat{H}(\bar{A},\bar{C},\bar{L})_{t-1}$: update $\bar{Y}_t^{\bar{d}_t}$ (or regress $\mathrm{offset}(\bar{Y}_t^{\bar{d}_t})$ + $\epsilon 1$ with weights $\hat{H}(\bar{A},\bar{C},\bar{L})_{t-1}$), with clever covariate (without censoring): $\hat{H}(\bar{A},\bar{L})_{t-1} = \prod_{s=0}^{t-1}\frac{\mathbb{1}(\bar{A}_s=\bar{d}_s)}{\widehat{\Pr}(A_s=d_s|\bar{A}_{s-1}=\bar{d}_{s-1},\bar{L}_s=\bar{l}_s)}$
4. $\hat{\psi}_T$ = mean of $\bar{Y}_1^{\bar{d}_1}$, get CI using influence curve

result is a $K+2$ multiply robust estimator (Díaz et al., 2021)

***t*argeted *m*inimum *l*oss-based *e*stimation**

target parameter $\Psi : \mathcal{M} \to \mathbb{R}$, with $\mathcal{M}$ the statistical model used

1. compute Ψ's pathwise derivative at P and its corresponding canonical gradient $D^*(P)$ (efficient influence curve)
2. define a loss $L()$ s. t. $P \to E_0 L(P)$ is minimized at true P_0
3. for a P in model $\mathcal{M}$ define a parametric working model $\{P(\epsilon) : \epsilon\}$ s. t. $P(\epsilon = 0) = P$ and a "score" $\frac{d}{d\epsilon}L(P(\epsilon))$ s. t. it (or linear combination of its components) equals $D^*(P)$ at P
4. compute $\epsilon_n^0 = \arg\min_\epsilon \sum_{i=1}^n L(P_n^0(\epsilon))(O_i)$, with initial estimate P_n^0, then first iteration $P_n^1 = P_n^0(\epsilon_n^0)$, repeat until $\epsilon_n^k = 0$
5. get TMLE estimate ψ_0 by plugging P_n^* into Ψ (substitution)
6. TMLE solves the efficient influence curve equation $\sum_{i=1}^n D^*(P_n^*)(O_i) = 0 \to$ asymptotic linearity and efficiency

can also be carried out for a relevant part Q instead of all of P

LMTP (Díaz et al., 2021) modified treatment policies

problems for (longitudinal) continuous or multi-valued A:

- fixed value counterfactuals unrealistic

- infinite-dimensional dose-response curve needs parametric assumptions or is not $n^{1/2}$ consistent
- positivity is often violated

solution: longitudinal MTP $A_t^{\mathrm{d}} = \mathrm{d}\left(A_t(\bar{A}_{t-1}^{\mathrm{d}}), H_t(\bar{A}_{t-1}^{\mathrm{d}})\right)$, e. g. threshold $(\max(c, a_t))$, shift ($a_t + \delta$ if positivity else a_t), stochastic (draw from $F(\mathrm{d}(A_t, H_t)|H_t)$; randomizer $\perp\!\!\!\perp U, P$), shifted propensity score (only for binary A)

identification for a given NPSEM, assumptions:

- *positivity* if (a_t, h_t) in $\mathrm{supp}\{A_t, H_t\}$ then $(\mathrm{d}(a_t, h_t), h_t)$ too
- *sequential randomization:*
  - *standard* $U_{A,t} \perp\!\!\!\perp \underline{U}_{L,t+1}|H_t$ (for stochastic LMTP)
  - *strong* $U_{A,t} \perp\!\!\!\perp (\underline{U}_{L,t+1}, \underline{U}_{A,t+1})|H_t$ (for other LMTP)

iterative process: set $m_{\tau+1} := Y$, for $t = \tau, ..., 1$:
$m_t : (a_t, h_t) \mapsto \mathrm{E}\left[m_{t+1}(A_{t+1}^{\mathrm{d}}, H_{t+1})|A_t = a_t, H_t = h_t\right]$
solve $\theta = \mathrm{E}\left[m_1(A_1^{\mathrm{d}}, L_1)\right]$

optimality *limitations:* threshold LMTPs can't be $n^{1/2}$ consistent as parameter not pathwise differentiable, continuous A can only be considered, if $\mathrm{d}(\cdot, h_t)$ *piecewise smooth invertible*

efficient influence curve (assumes $\mathrm{d} \perp\!\!\!\perp \mathrm{P}$):

$$EIF\left(\mathrm{E}\left[m_1(A^{\mathrm{d}}, L_1)\right]\right) = \phi_1(Z) - \theta$$

with $r_t(a_t, h_t) = \frac{g_t^{\mathrm{d}}(a_t|h_t)}{g_t(a_t|h_t)}$ and $\phi_t : z \mapsto \sum_{s=t}^{\tau}\left(\prod_{k=t}^{s} r_k(a_k, h_k)\right)$ $\left\{m_{s+1}(a_{s+1}^{\mathrm{d}}, h_{s+1}) - m_s(a_s, h_s)\right\} + m_t(a_t^{\mathrm{d}}, h_t)$

estimation use Super Learner for $\hat{r}_t$ and $\hat{m}_t$

- ***g-methods:*** asymptotically linear and $n^{1/2}$ consistent if models correctly specified, asymptotic distribution generally unknown

substitution (standardization): $\hat{\theta}_{\mathrm{sub}} = \frac{1}{n}\sum_{i=1}^{n} \hat{m}_1(A_{1,i}^{\mathrm{d}}, L_{1,i})$
IPTW: $\hat{\theta}_{\mathrm{iptw}} = \frac{1}{n}\sum_{i=1}^{n}\left(\prod_{t=1}^{\tau} \hat{r}_t(A_{t,i}, H_{t,i})\right) Y_i$

- ***TMLE:*** use sample splitting and cross-fitting with sets $\mathcal{T}_j$, TMLE sets cross-validated EIF $P_n\left\{\phi_1(., \tilde{\eta}_j(.)) - \hat{\theta}_{\mathrm{tmle}}\right\}$ to zero

τ+1 multiply robust & $n^{1/2}$ consistent (if nuisance consistant)

step 1: initialize $\tilde{\eta} = \hat{\eta}$ and $\tilde{m}_{\tau+1,j(i)}(A_{\tau+1,i}^{\mathrm{d}}, H_{\tau+1,i}) = Y_i$
step 2: compute τ weights $\omega_{s,i} = \prod_{k=1}^{s} \hat{r}_{k,j(i)}(A_{k,i}, H_{k,i})$
step 3: for $t = \tau, ..., 1$: fit generalized linear tilting model

$$\mathrm{link}\, \tilde{m}_t^{\epsilon}(A_{t,i}, H_{t,i}) = \epsilon + \mathrm{link}\, \tilde{m}_{t,j(i)}(A_{t,i}, H_{t,i})$$

with the canonical link and use $\hat{\epsilon}$ to update $\tilde{m}_{t,j(i)}^{\hat{\epsilon}}$
step 4: $\hat{\theta}_{\mathrm{tmle}} = \frac{1}{n}\sum_{i=1}^{n} \tilde{m}_{1,j(i)}(A_{1,i}^{\mathrm{d}}, L_{1,i})$

- ***SDR:*** 2^{τ} multiply robust (sequentially double robust) and same rate of $n^{1/2}$ consistency as TMLE, better finite sample behavior than TMLE but estimate is not guaranteed to be in support

step 0: cross-fit estimates $\hat{r}_{1,j(i)}, ..., \hat{r}_{\tau,j(i)}$
step 1: $\phi_{\tau+1}(Z_i; \breve{\eta}_{\tau,j(i)}) = Y_i$
step 2: for $t = \tau, ..., 1$:

- compute pseudo-outcome $\breve{Y}_{t+1,i} = \phi_{t+1}(Z_i; \breve{\eta}_{\tau,j(i)})$
- for $j = 1, ..., J$: regress $\breve{Y}_{t+1,i}$ on $(A_{t,i}, H_{t,i})$ only using $i \in \mathcal{T}_j$, with $\breve{m}_{t,j}$ output, update $\underline{\breve{\eta}}_{t,j} = (\hat{r}_{t,j}, \breve{m}_{t,j}, ..., \hat{r}_{\tau,j}, \breve{m}_{\tau,j})$

step 3: $\hat{\theta}_{\mathrm{sdr}} = \frac{1}{n}\sum_{i=1}^{n} \phi_1(Z_i, \breve{\eta}_{j(i)})$

∗ ***estimate density ratio r_t:*** duplicate dataset, where duplicates get assigned A_t^{d} with indicator $\Lambda \in \{0, 1\}$
$r_t(a_t, h_t) \overset{1}{=} \frac{p^{\lambda}(a_t, h_t|\Lambda=1)}{p^{\lambda}(a_t, h_t|\Lambda=0)} \overset{2}{=} \frac{P^{\lambda}(\Lambda=1|A_t=a_t, H_t=h_t)}{P^{\lambda}(\Lambda=0|A_t=a_t, H_t=h_t)} \overset{3}{=} \frac{u_t^{\lambda}(a_t, h_t)}{1-u_t^{\lambda}(a_t, h_t)}$
with 1 definition of r_t, 2 Bayes rule, and 3 by definition
⇒ any classification method can be used (e. g. Super Learning), cross-fitting should be used

Methods for continuous A (Kennedy et al., 2017)

doubly robust methods possible for continuous A for *parametric* effect curves otherwise a $\sqrt{n}$ consistent estimator can not exist

procedure: found double robust ξ using efficient influence curve in $\mathbb{E}\left\{\xi(Z; \bar{\pi}, \bar{\mu})|A = a\right\} = \theta(a)$, with data Z and nuisance π, μ

step 1: estimate nuisance π, μ and predict
step 2: construct pseudo-outcome $\hat{\xi}(Z; \hat{\pi}, \hat{\mu})$ and regress on A (e. g. using local linear kernel regression)

consistent if: either $\bar{\pi} = \pi$ or $\bar{\mu} = \mu$, $\theta(a)$ twice continuously differentiable (and two other items are continuous, assumptions on the kernel part and the function class of nuisance)

asymptotic normality if at least one nuisance is fast enough

TMLE version a clever covariate is given by the authors

References

If no citation is given, the information is taken from the book (Hernán and Robins, 2020)

Abbott, S. (2015). *Understanding analysis*. Springer.

Díaz, I., Williams, N., Hoffman, K. L., and Schenck, E. J. (2021). Nonparametric causal effects based on longitudinal modified treatment policies. *Journal of the American Statistical Association*, pages 1–16.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the american statistical association*, 69(346):383–393.

Hernán, M. A. and Robins, J. M. (2020). *Causal inference: what if*. Boca Raton: Chapman & Hall/CRC.

Hernán, M. A. and Robins, J. M. (2023). *Causal inference: what if*. Boca Raton: Chapman & Hall/CRC.

Kennedy, E. H., Ma, Z., McHugh, M. D., and Small, D. S. (2017). Non-parametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(4):1229–1245.

Pearl, J. (2019). The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, 62(3):54–60.

Petersen, M. L. and van der Laan, M. J. (2014). Causal models and learning from data: integrating causal modeling and statistical estimation. *Epidemiology (Cambridge, Mass.)*, 25(3):418–426.

Schomaker, M., Luque-Fernandez, M. A., Leroy, V., and Davies, M.-A. (2019). Using longitudinal targeted maximum likelihood estimation in complex settings with dynamic interventions. *Statistics in medicine*, 38(24):4888–4911. ISBN: 0277-6715 Publisher: Wiley Online Library.

van der Laan, M. J. and Gruber, S. (2012). Targeted minimum loss based estimation of causal effects of multiple time point interventions. *The international journal of biostatistics*, 8(1).

van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007). Super learner. *Statistical applications in genetics and molecular biology*, 6(1). Article 24.

van der Laan, M. J., Rose, S., et al. (2011). *Targeted learning: causal inference for observational and experimental data*, volume 4. Springer.

van der Laan, M. J. and Rubin, D. (2006). Targeted maximum likelihood learning. *The international journal of biostatistics*, 2(11).

Young, J. G., Cain, L. E., Robins, J. M., O'Reilly, E. J., and Hernán, M. A. (2011). Comparative effectiveness of dynamic treatment regimes: an application of the parametric g-formula. *Statistics in biosciences*, 3:119–143.

Zepeda-Tello, R., Schomaker, M., Maringe, C., Smith, M. J., Belot, A., Rachet, B., Schnitzer, M. E., and Luque-Fernandez, M. A. (2022). The delta-method and influence function in medical statistics: a reproducible tutorial. *arXiv preprint arXiv:2206.15310*.