
Causal Inference

a summary

Contents

1	General	3		
	Causal Roadmap	3		
	Average Causal Effect	3		
	Target Trial	3		
	Identifiability Conditions	3		
	Effect Modification	3		
	Interaction	4		
	NPSEM	4		
	Causal DAG	4		
	Noncausal DAGs	4		
	SWIGs	4		
	Confounding	4		
	Selection Bias	4		
	Measurement Bias	4		
	Random Variability	5		
	Time-Varying Treatments	5		
	Sequential Identifiability	5		
	Treatment-Confounder Feedback	5		
	Causal Mediation	5		
2	Models	6		
	Modeling	6		
	Variable Selection	6		
	Super Learning	6		
			Marginal Structural Models	6
	2.1	Traditional Methods	6	
		Stratification	6	
		Outcome Regression	6	
		Propensity Score Methods	6	
		Instrumental Variable Estimation	6	
		Causal Survival Analysis	6	
	2.2	G-Methods	7	
		G-Methods	7	
		Standardization	7	
		IP Weighting	7	
		G-Estimation	8	
	2.3	Doubly Robust Methods	8	
		Double-Robustness	8	
		Machine Learning	8	
		Advantages	8	
		Influence Curve	8	
		Simple Plug-In Estimator	9	
		Augmented IP Weighting (AIPTW)	9	
		TMLE	9	
		LTMLE	9	
		TMLE advanced	9	
		AIPTW	10	

1 General

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 assumptions (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 $E[Y^{a=1}] \neq E[Y^{a=0}]$

$$E[Y^a] = \sum_y yp_{Y^a}(y) \quad (\text{discrete})$$

$$= \int y f_{Y^a}(y) dy \quad (\text{continuous})$$

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: $-1/\text{risk difference}$

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 time, outcome \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:* $Y^a \perp\!\!\!\perp A \hat{=}$ randomized experiment, counterfactuals are missing completely at random (MCAR)
- *conditional:* $Y^a \perp\!\!\!\perp A|L \hat{=}$ 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 $\xrightarrow{\text{may}}$ misspecification bias

Effect Modification A on Y varies across levels of V

null average causal effect \neq 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

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

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 nonparametric structural equation model

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

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 \cap FCISTG \Rightarrow 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 \Rightarrow FFRCISTG (assume DAGs represent latter)

NPSEM-IE assume crossworld independencies \rightarrow 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 \Rightarrow independence

under faithfulness: independence \Rightarrow 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 $L \rightarrow Y$)?

if Y^l is not well-defined, but $L \rightarrow Y$, then the graph is not causal

statistical interpretation: only $A \rightarrow Y$ 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 single world intervention graphs

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 $L \rightarrow Y^a \perp\!\!\!\perp A|L$ (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 $L \Rightarrow$ 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 : $\Pr[Y^a = 1] = \sum_m \Pr[M = m|A = a] \sum_{a'} \Pr[Y = 1|M = m, A = a'] \Pr[A = a']$

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: $\Pr[Y=0|E=e, A=a] = g(e)h(a) \rightarrow$ no interaction between E and A on the multiplicative scale; if $Y = 0$ is conditionally independent, then $Y = 1$ can't be as $\Pr[Y=1|E=e, A=a] = 1 - g(e)h(a) \rightarrow$ 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 \rightarrow bias, if: $A \rightarrow Y$ or dependent or differential

reverse causation bias caused by e.g. recall bias: independent but differential A (caused by $Y \rightarrow U_A$)

misclassified treatment: assignment Z does not determine A
exclusion restriction: ensure $Z \not\rightarrow Y$, e.g. via double-blinding

- **per-protocol effect**: either as-treated (\rightarrow confounded) or restricted to protocol adhering individuals (\rightarrow selection bias)
- **intention-to-treat effect** (\rightarrow 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 is of interest (Hernán and Robins, 2023)

Random Variability 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 : $\bar{A}_k = (A_0, A_1, \dots, A_k)$

shorthand: always treated $\bar{A} = \bar{1}$, never treated $\bar{A} = (\bar{0})$

static strategy: $g = [g_0(\bar{a}_{-1}), \dots, g_K(\bar{a}_{K-1})]$

dynamic strategy: $g = [g_0(\bar{l}_0), \dots, g_K(\bar{l}_K)]$

stochastic strategy: non-deterministic g

optimal strategy is where $E[Y^g]$ is maximized (if high is good)

Sequential Identifiability sequential versions of

exchangeability: $Y^g \perp\!\!\!\perp A_k | \bar{A}_{k-1} \quad \forall g, k = 0, 1, \dots, K$

conditional exchangeability:

$$(Y^g, L_{k+1}^g) \perp\!\!\!\perp A_k | \bar{A}_{k-1} = g(\bar{L}_k), \bar{L}^k \quad \forall g, k = 0, 1, \dots, K$$

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

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

consistency:

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

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

generalized backdoor criterion (static strategy): all backdoors into A_k (except through future treatment) are blocked $\forall k$

static sequential exchangeability for $Y^{\bar{a}}$

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

use SWIGs to visually check d-separation

time-varying confounding $E[Y^{\bar{a}} | L_0] \neq E[Y | A = \bar{a}, L_0]$

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
 $Y^g = Y \quad \forall g \Rightarrow Y \perp\!\!\!\perp A_0 | L_0$ & $Y \perp\!\!\!\perp A_1 | A_0, L_0, L_1$; therefore:
if last two independences don't hold, one assumption is violated
g-null theorem: $E[Y^g] = E[Y]$, if the two independences hold
(\Rightarrow sharp null: only if strong faithfulness (no effect cancelling))

Causal Mediation (Hernán and Robins, 2023)

$A \xrightarrow{M} Y$ seen as longitudinal with k_0 : A and k_1 : M
decomposition of $E[Y^{a=1}] - E[Y^{a=0}]$ into cross-world quantities

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

$$E[Y^{a=1, M^{a=0}}] - E[Y^{a=0, M^{a=0}}]$$

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

$$E[Y^{a=1, M^{a=1}}] - E[Y^{a=1, M^{a=0}}]$$

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

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

interventional interpretation

additional assumptions:

$$A \xrightarrow{N} O \xrightarrow{M} Y$$

thick arrows are deterministic

two assumptions: no $N \rightarrow Y$ and no $O \rightarrow M$ ("no controlled direct effect" assumptions)

2 Models

Modeling data are a sample from the target population

estimand: quantity of interest, e. g. $E[Y|A = a]$
estimator: function to use, e. g. $\hat{E}[Y|A = a]$
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 \rightarrow true value)

parsimonious model: few parameters estimate many quantities

bias-variance trade-off:

wiggleness $\uparrow \rightarrow$ misspecification bias \downarrow , CI width \uparrow

Variable Selection can induce bias if L includes:

(descendant 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

$\text{logit}(Y = 1|Z) = \sum_i \alpha_i Z_i$, 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

associational model $E[Y|A] =$ causal model $E[Y^a]$

step 1: estimate/model $f[A|L]$ (and $f[A]$) \rightarrow 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

2.1 Traditional Methods

Stratification calculate risk for each stratum of L

only feasible if enough data per stratum

Outcome Regression often assume no effect modification

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

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*

P 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 (\rightarrow 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) \rightarrow weak instrument)
2. **exclusion restriction**: Z affects Y at most through A
 - (a) population level: $E[Y^{z,a}] = E[Y^{z',a}]$ (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 $A \rightarrow Y \forall i$ (ii) constant average effect $A \rightarrow Y \forall A$ (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) $\xrightarrow{\text{data}}$ (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 \div measure of compliance

(1,2b,3a,4a): ATE; (1,2b,3a,4b): ATE in compliers

binary Z : $\frac{E[Y|Z=1] - E[Y|Z=0]}{E[A|Z=1] - E[A|Z=0]}$, continuous Z : $\frac{Cov(Y,Z)}{Cov(A,Z)}$;
 can be calculated as *two-stage-least-squares estimator*:

1. $E[A|Z]$ 2. $E[Y|Z] = \beta_0 + \beta_1 \hat{E}[A|Z]$ 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 $L \rightarrow$ 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 ?) \rightarrow 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** SW^A 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 weights** SW^A in structural marginal model
 $\text{logit } \Pr[D_{k+1}^{a,c=0} = 0|D_k^{a,c=0} = 0] = \beta_{0,k} + \beta_{1,a} + \beta_{2,a}k$
 - **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):** $\frac{\Pr[D_k^c=1|L,A]}{\Pr[D_k^c=0=1|L,A]} = \exp[\gamma_k(L,A;\psi)]$ (log-linear has no upper limit $1 \rightarrow$ rare failure \uparrow ; if \downarrow , 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)

2.2 G-Methods

G-Methods generalized treatment contrasts: adjust for (surrogate) confounders L

- **standardization:** two types of g-formula
- **IP weighting:** 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 \rightarrow 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 $E[Y^{a,c=0}]$
standardization $E[Y|A=a] = \int E[Y|L=l, A=a, C=0] dF_L[l]$

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 \rightarrow use IP weights

time-varying censoring \bar{C} : monotonic type of missing data

standardization: $\int f(y|\bar{a}, \bar{c}=\bar{0}, \bar{l}) \prod_{k=0}^K dF(l_k|\bar{a}_{k-1}, c_{k-1}=0, \bar{l}_{k-1})$

IP weighting:

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

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

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

weighted average of stratum-specific risks; unknowns can be estimated non-parametrically or modeled

no need to estimate $f_L[l]$ /integrate as empirical distribution can be used: estimate outcome model \rightarrow predict counterfactuals on whole dataset \rightarrow average the results (\rightarrow CI by bootstrapping)

for discrete L $E[Y|A=a] = \sum_l E[Y|L=l, A=a] \Pr[L=l]$

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

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

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

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

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

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

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

estimation (Young et al., 2011; Schomaker et al., 2019)

1. model $f(l_k|\bar{a}_{k-1}, \bar{l}_{k-1})$ and $E[Y|\bar{A}=\bar{a}, \bar{L}=\bar{l}]$
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)

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

estimation (Schomaker et al., 2019)

1. model inside out: $Q_T = E[Y_T|\bar{A}_{T-1}, \bar{L}_T]$ to $Q_0 = E[Q_1|\bar{L}_0]$, 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 $L_0 \rightarrow A_0 \rightarrow Y_0 \rightarrow L_1 \rightarrow A_1 \rightarrow Y_1$, $\bar{a} = (a_0, a_1)$

$$E[Y_1^{\bar{a}}] \stackrel{\text{CE}}{=} E[E[Y_1^{\bar{a}}|A_0=a_0, L_0]]$$

$$(\text{ice}) \stackrel{\text{CE}^*}{=} E[E[E[Y_1|\bar{L}, \bar{A}=\bar{a}, Y_0] | A_0=a_0, L_0]]$$

$$\stackrel{\text{LTP}}{=} E\left[\sum_{l_1} E[Y_1|A_0=a_0, \bar{L}, Y_0] \Pr[l_1|a_0, l_0, y_0]\right]$$

$$\stackrel{\text{LTP}}{=} \sum_{l_0} \left[\sum_{l_1} E[Y_1|A_0=a_0, \bar{L}, Y_0] \Pr[l_1|a_0, l_0, y_0] \right] \Pr[l_0]$$

$$(\text{jde}) \stackrel{\text{sum}}{=} \sum_{\bar{l}} E[Y_1|A_0=a_0, \bar{L}, Y_0] \Pr[l_1|a_0, l_0] \Pr[l_0]$$

CE: conditional expectation; *: exchangeability;

LTP: law of total probability

IP Weighting inverse probability of treatment (g-formula)

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

unknowns can be estimated non-parametrically or modeled

pseudo-population: everyone is treated & untreated ($L \not\rightarrow A$)

FRCISTG (fully randomized causally interpreted structured graph): probability tree for $L \rightarrow A \rightarrow Y$, can be used to calculate/visualize simulation of values for A

for discrete A, L $f[a|l] = \Pr[A=a, L=l]$

estimators: Horvitz-Thompson; Hajek (modified version)

stabilized weights SW^A should have an average of 1 (check!)

\rightarrow pseudo-population same size \rightarrow CI width \downarrow

time-varying

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

G-Estimation (additive) structural nested models

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

$$H(\psi^\dagger) = Y - \psi_1 A$$

find ψ^\dagger which renders $\alpha_1 = 0$; 95%-CI: all ψ^\dagger for which $p > 0.05$
closed-form solution for linear models

derivation: $H(\psi^\dagger) = Y^{a=0}$

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

$Y^{a=0}$ 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

$$\text{additive: } E[Y^a - Y^{a=0} | A = a, L] = \beta_1 a + \beta_2 a L$$

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

multiplicative is preferred if Y always positive, but does not extend to longitudinal case

semi-parametric: agnostic about β_0 and effect of $L \rightarrow$ robust \uparrow

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 $\alpha_1 \rightarrow$ plot α_1 vs. $\psi^\dagger \rightarrow$ 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

structural nested mean models separate effect of each a_k

$$E[Y^{\bar{a}_{k-1}, a_k, \bar{0}_{k+1}} - Y^{\bar{a}_{k-1}, \bar{0}_{k+1}} | \bar{L}^{\bar{a}_{k-1}} = \bar{L}_k, \bar{A}_{k-1} = \bar{a}_{k-1}] =$$

$$a_k \gamma_k(\bar{a}_{k-1}, \bar{L}_k, \beta)$$

calculations

$$H_k(\psi^\dagger) = Y - \sum_{j=k}^K A_j \gamma_j(\bar{A}_{j-1}, \bar{L}_j, \psi^\dagger)$$

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

$$\text{logit Pr} [A_k = 1 | H_k(\psi^\dagger), \bar{L}_k, \bar{A}_{k-1}] =$$

$$\alpha_0 + \alpha_1 H_k(\psi^\dagger) + \alpha_2 w_k(\bar{L}_k, \bar{A}_{k-1})$$

find α_1 that is closest to zero

a closed form estimator exists for the linear case

2.3 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 $\frac{1}{\pi(l)} - \frac{1}{\pi(l)}$ and $b(l) - \hat{b}(l)$ known as second order bias

time-varying: multiple robustness for $k = 0, 1, \dots, K$

$K+2$ 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 \rightarrow 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 $E[\hat{\psi} - \psi | T_r] / se(\hat{\psi})$ converges to $0 \rightarrow \hat{\psi}$ with sample splitting is asymptotically normal and unbiased \rightarrow 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):

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

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 (Hampel, 1974; Van der Laan et al., 2011; Zepeda-Tello et al., 2022) how robust is an estimator?

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

for estimator T and distribution P_n with $0 < \epsilon < 1$

can also be rewritten as

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

last part is a **directional derivative** at P_n (empirical probability measure that puts mass $1/n$ on O_i) in direction $(\delta_O - P_n)$

$\bar{IC}(P_0) = 0$ and $\text{Var}(IC(P_0))$ is the asymptotic variance of the standard estimator $\sqrt{n}(\psi_n - \psi_0)$, therefore $\text{Var}(\hat{\psi}(P_n)) = \frac{\text{Var}_{IC}}{n}$

efficient IC: an estimator is asymptotically efficient \Leftrightarrow its influence curve is the efficient influence curve $IC(O) = D^*(O)$ estimand written as a function of θ , i.e. $\psi := \phi(\theta)$, we know the variance of $\hat{\theta}$, but what is the variance of $\hat{\psi}$

classical delta method: (under regularity conditions)

distribution of $\phi(\hat{\theta})$ can be approximated as a normal with a variance proportional to ϕ 's rate of change at θ :

$$\sqrt{n}(\phi(\hat{\theta}_n) - \phi(\mu)) \stackrel{\text{approx}}{\sim} N(0, \phi'(\theta)\sigma^2)$$

which allows for Wald-type CI: $\hat{\theta}_n \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\phi'(\theta)\sigma^2}{n}}$ (very similar to central limit theorem)

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[†]): $\phi(\hat{\theta}_n) \approx \phi(\theta) + \partial_{v:=\hat{\theta}-\theta} \phi(\theta)$
- classical delta method**: if $\{r_n\}_{n=1}^\infty$ with $\lim_{n \rightarrow \infty} r_n = \infty$, where $r_n(\hat{\theta}_n - \theta)$ converges to $Z \sim N(0, 1)$ (e.g. $r_n = \sqrt{n/\sigma^2}$), then

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

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

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

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

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

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

use delta: $\phi(\mathbb{P}_X) \approx \phi(\mathbb{P}_X) + \text{IF}_{\phi, \mathbb{P}_X}(Y)$, if $(\hat{\theta}_n - \theta) \stackrel{n \rightarrow \infty}{\rightsquigarrow} N(\cdot, \cdot)$

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

where $\widehat{\text{Var}}[\text{IF}_{\phi, \mathbb{P}_X}(Y)] = \frac{1}{n} \sum_{i=1}^n (\text{IF}_{\phi, \mathbb{P}_X}(X_i))^2$, 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

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 $\hat{f}(A_m | \bar{A}_{m-1}, \bar{L}_m)$ (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 $\hat{T}_{K+1} := Y$, recursively for $m = K, K-1, \dots, 0$:
 - (a) fit outcome regression on \hat{T}_{m+1} with variable $\widehat{W}^{\bar{A}_m}$
 - (b) calculate \hat{T}_m using the outcome model with $\widehat{W}^{\bar{A}_{m-1}, a_m}$
3. calculate standardized mean outcome $\hat{E}[Y^a] = E[\hat{T}_0]$

Augmented IP Weighting (AIPTW) (Hernán and Robins, 2023)

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

Relationship between AIPTW and TMLE for causal effect:

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

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

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

TMLE (Van der Laan et al., 2011)

targeted maximum likelihood estimation

$$O = (W, A, Y) \sim P_0$$

target $\Psi(P_0) = \Psi(\bar{Q}_0, Q_{W,0}) = \psi_0$,

often: $E_{W,0} [E_0(Y|A=1, W) - E_0(Y|A=0, W)]$

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 $f(Y|A, W)$ needs to be estimated, just the relevant portion, *typically average outcome* $E_0(Y|A, W) \rightarrow \text{efficiency} \uparrow$

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

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: $\text{logit} \bar{Q}_n^1(A, W) = \text{logit} \bar{Q}_n^0(A, W) + \epsilon_n H_n^*(A, W)$ (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.: $\psi_n^{\text{TMLE}} = \frac{1}{n} \sum_{i=1}^n [\bar{Q}_n^1(1, W_i) - \bar{Q}_n^1(0, W_i)]$

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), substitution estimator (respects global constraints \rightarrow more robust to outliers and sparsity)

closed form inference based on the influence curve:

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

$$+ \underbrace{\left[\bar{Q}_n^1(1, W_i) - \bar{Q}_n^1(0, W_i) - \psi_{\text{TMLE}, n} \right]}_b$$

TMLE sets the mean of the IC, \overline{IC}_n , to zero (b has already mean zero, see third step, the first part of a is the clever covariate)

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

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

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

LTMLE longitudinal

for $t = T, \dots, 1$:

1. model $E(Y_t | \bar{A}_{t-1}, \bar{L}_t)$ (fit on individuals that are uncensored and alive at $t-1$)
2. plug in $\bar{a}_{t-1} = \bar{d}_{t-1}$; use regression from 1 to predict outcome at time t , ie. $\bar{Y}_t^{\bar{d}_t}$
3. update estimate with $Y_t = \text{offset}(\text{step2resultint}) + \epsilon \times \text{clevercovariate}$: predict $\bar{Y}_t^{\bar{d}_t}$ (alternatively the clever covariate can be used as a weight)
4. $\hat{\psi}_T = \text{mean of } \bar{Y}_1^{\bar{d}_1}$

TMLE advanced (Van der Laan et al., 2011)

targeted minimum loss-based estimation

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

1. compute its pathwise derivative at P and corresponding canonical gradient $D^*(P)$ (efficient influence curve: a function of O with mean zero under P)
2. define loss function $L()$ s.t. $P \rightarrow E_0 L(P)$ is minimized at true P_0 (or just relevant Q)
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” $\frac{d}{d\epsilon} L(P(\epsilon))$: score (or linear combination of its components) equals $D^*(P)$ at P (or just relevant Q)
4. with initial estimate P_n^0 , compute

$\epsilon_n^0 = \arg \min_{\epsilon} \sum_{i=1}^n L(P_n^0(\epsilon))(O_i)$, calculate first iteration $P_n^1 = P_n^0(\epsilon_n^0)$, repeat until $\epsilon_n^k = 0$ (or just relevant Q)

5. get TMLE estimate ψ_0 as the substitution estimator
pluggint P_n^* into Ψ
6. TMLE solves the efficient influence curve equation
 $0 = \sum_{i=1}^n D^*(P_n^*)(O_i) \rightarrow$ esymptotic linearity and efficiency

$\mathcal{L}(O) = \overbrace{\Pr(Y|A, W)}^{Q_Y} \overbrace{\Pr(A|W)}^g \overbrace{\Pr(W)}^{Q_W}$: g itself is not needed as we intervene on treatment, but it can help improving the estimate of Q_Y

$H(A, W)$ depends on target parameter and loss function but is a function of the propensitiy score update initial fit
 $\bar{Q}_n^* = \bar{Q}_n^0 + \hat{\epsilon}H(A, W)$

valid inference, good finite sample performance,

$H(A, W)$ comes from the influence curve, targeting ensures mean of efficient influence curve $D^*(P)$ is zero

TMLE solves $P_n D^*(P_n^*) = 0$

TMLE is a substitution estimator

$\psi_n^{TMLE} = \frac{1}{2} \sum_{i=1}^n \bar{Q}_n^*(1, W_i) - \frac{1}{2} \sum_{i=1}^n \bar{Q}_n^*(0, W_i)$ therefore mean of b is zero

targeting step makes sure a also has mean zero

MLE solves $\sum_{i=1}^n H(A_i, W_i) [Y_i - \bar{Q}_n^*(A_i, W_i)] = 0$ where
 $\bar{Q}_n^*(A_i, W_i) = \hat{\epsilon}H(A, W) + \bar{Q}_n^0$ therefore obvious choice:

$$H(A, W) = \frac{A}{g(1, W)} - \frac{1-A}{g(0, W)}$$

influence curve based inference: asymptotic linearity

$$\sqrt{n} (\psi_n^{TMLE} - \psi_0) \xrightarrow{D} N(0, \sigma^2)$$

AIPW augmented inverse probability of treatment weighting

disadvantages (Van der Laan et al., 2011): ignores global constraints \rightarrow often unstable under sparsity, sometimes not well-defined

References

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

- 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.
- 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., 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.
- 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*.