Causal Inference

a summary

Contents

1 General **3**

2 Models **5**

2.1 Traditional Methods 5

2.2 G-Methods 5

2.2.1 Time-varying A 6

2.3 Doubly Robust Methods 7

3 Longitudinal Data **9**

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

$$\mathrm{E}\left[Y^{a}\right] = \sum_y y p_{Y^a}(y) \qquad \text{(discrete)}$$

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

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: $\Pr\left[Y^{a=1} = 1\right] - \Pr\left[Y^{a=0} = 1\right]$

risk ratio: $\frac{\Pr\left[Y^{a=1}=1\right]}{\Pr\left[Y^{a=0}=1\right]}$

odds ratio: $\frac{\Pr\left[Y^{a=1}=1\right]/\Pr\left[Y^{a=1}=0\right]}{\Pr\left[Y^{a=0}=1\right]/\Pr\left[Y^{a=0}=0\right]}$

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 $E\left[Y|A=1\right]$ and $E\left[Y|A=0\right]$, **causation** compares $E\left[Y^{a=1}\right]$ and $E\left[Y^{a=0}\right]$ (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 $\Pr\left[A = a|L = l\right] > 0 \ \forall\, l$ with $\Pr\left[L = l\right] > 0$;

$$f_L(l) \neq 0 \Rightarrow f_{A|L}(a|l) > 0 \ \forall a, l$$

- 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: $\Pr\left[A = 1|Y^{a=0}, L\right] = \Pr\left[A = 1|L\right]$

additional conditions:

correct measurement mismeasurement of A, Y, L results in bias

correct model specification models $\overset{\text{may}}{\to}$ 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*on*p*aramentric *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 treee 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

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

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: $\Pr[Y{=}0|E{=}e, A{=}a] = g(e)h(a)$ $\rightarrow$ 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* differantial

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)

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: also condition on approximate ancillaries (continuity principle)

curse of dimensionality: difficult to do conditionality principle

2 Models

Modeling

data are a sample from the target population

estimand:	quantity of interest,	e. g. $\mathrm{E}[Y \mid A=a]$
estimator:	function to use,	e. g. $\widehat{\mathrm{E}}[Y \mid A=a]$
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)

Machine Learning

L is high-dimensional
use lasso or ML for IP weighting/standardization
but: ML does not guarantee elimination of confounding and has largely unknown statistical properties
→ **doubly robust estimator:** consistent if bias $< \frac{1}{\sqrt{n}}$
sample splitting: train estimators on training sample, 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
problems: unclear choice of algorithm, is bias small enough?

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

$$\operatorname{logit}(Y=1 \mid Z) = \sum_i \alpha_i Z_i, \text{ with } 0 < \alpha_i < 1 \text{ and } \sum \alpha_i = 1$$

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)

2.1 Traditional Methods

Stratification

calculate risk for each stratum of L

only feasible if enough data per stratum

Outcome regression chapter 15

instrumental variable estimation chapter 16

causal survival analysis chapter 17 (and technical point 22.3)

2.2 G-Methods

G-Methods

*g*eneralized 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

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

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

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 → predict counterfactuals on whole dataset → average the results (→ CI by bootstrapping)

for discrete L $\mathrm{E}[Y \mid A=a] = \sum_l \mathrm{E}[Y \mid L=l, A=a] \Pr[L=l]$

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)

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

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

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 $\hat{Y}^{\bar{a}}_{K,i}$ (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 $\hat{Q}^{\bar{a}}_{0,i}$ (bootstrap for CI)

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

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

$$\mathrm{E}\left[Y_1^{\bar{a}}\right] \overset{\text{CE}}{=} \mathrm{E}\left[\mathrm{E}\left[Y_1^{\bar{a}}|A_0{=}a_0, L_0\right]\right]$$

$$\text{(ice)} \quad \overset{\text{CE*}}{=} \mathrm{E}\left[\mathrm{E}\left[\mathrm{E}\left[Y_1|\bar{L}, \bar{A}{=}\bar{a}, Y_0\right]|A_0{=}a_0, L_0\right]\right]$$

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

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

$$\text{(jde)} \quad \overset{\text{sum}}{=} \sum\nolimits_{\bar{l}} \mathrm{E}\left[Y_1|A_0{=}a_0, \bar{L}, Y_0\right] \Pr\left[l_1|a_0, l_0\right] \Pr\left[l_0\right]$$

CE: conditional expectation; *: exchangeability;
LTP: law of total probability

Marginal Structural Models

association is causation in the IP weighted pseudo-population

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

step 1: estimate/model $f\left[A|L\right]$ (and $f\left[A\right]$) → get $(S)W^A$

step 2: estimate regression parameters for pseudo-population

effect modification variables V can be included (e. g. $\beta_0 + \beta_1 a + \beta_2 V a + \beta_3 V$; technically not marginal anymore), $SW^A(V) = \frac{f[A|V]}{f[A|L]}$ more efficient than SW^A

Censoring

measuring joint effect of A and C

$$\mathrm{E}\left[Y^{a,c=0}\right] \text{ is of interest}$$

standardization $\mathrm{E}\left[Y|A=a\right] = \int \mathrm{E}\left[Y|L{=}l, A{=}a, C{=}0\right] dF_L\left[l\right]$

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 can only adjust for confounding, not selection bias → use IP weights

G-Estimation

(additive) structural nested models

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

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

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

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

$Y^{a=0}$ 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

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

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

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

semi-parametric: agnostic about β_0 and effect of L → 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

IP Weighting

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

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

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 $\boldsymbol{A}, \boldsymbol{L}$ $f\left[a|l\right] = \Pr\left[A=a, L=l\right]$

estimators: Horvitz-Thompson; Hajek (modified version)

stabilized weights $\boldsymbol{SW^A}$ should have an average of 1 (check!) → pseudo-population same size → CI width ↓

Standardization and IP Weighting

are equivalent, ***but*** if modeled, different "no misspecification" assumptions:

standardization: outcome model

IP weighting: treatment model

doubly robust estimators: reduce model misspecification bias, consistent if either model is correct; ***e. g.:***

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

2.2.1 Time-varying A

IP Weighting

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

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

Doubly Robust Estimator

sequential estimation

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 $\widehat{\mathrm{E}}\left[Y^{\bar{a}}\right]=\mathrm{E}\left[\widehat{T}_0\right]$

valid, if treatment or outcome model correct, or treatment correct until k and outcome otherwise ($k+1$ robustness)

G-Estimation

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

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

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

find α_1 that is closest to zero

a closed form estimator exists for the linear case

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:

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

2.3 Doubly Robust Methods

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

(Van der Laan et al., 2011)

influence curve IC(O): how robust is estimator toward extreme values

$meanIC(P_0)=0$ and its finite variance is the asymptotic variance of the standard estimator

TMLE

(Van der Laan et al., 2011)

*t*argeted *m*aximum *l*ikelihood *e*stimation

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

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

often: $\mathrm{E}_{W,0}\left[\mathrm{E}_0(Y|A{=}1,W)-\mathrm{E}_0(Y|A{=}0,W)\right]$

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* $\mathrm{E}_0(Y|A,W)$ → efficiency ↑

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

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: $\operatorname{logit}\bar{Q}_n^1(A,W)=\operatorname{logit}\bar{Q}_n^0(A,W)+\epsilon_n H_n^*(A,W)$ (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)

closed form inference based on the influence curve:

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

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

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

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

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: $2\left[1-\Phi\left(\left|\frac{\psi_{TMLE,n}}{\sigma_n/\sqrt{n}}\right|\right)\right]$

TMLE advanced

(Van der Laan et al., 2011)

*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 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. $P\to E_0L(P)$ 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 olves the efficient influence curve equation $0 = \sum_{i=1}^{n} D^*(P_n^*)(O_i)$ → esymptotic linearity and efficiency

---

$\mathcal{L}(O) = \overbrace{\Pr(Y|A,W)}^{Q_Y}\overbrace{\Pr(A|W)}^{g}\overbrace{\Pr(W)}^{Q_W}$

$H(A,W)$ 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 $\sum_{i=1}^{n} H(A_i,W_i)\left[Y_i - \bar{Q}_n^*(A_i,W_i)\right] = 0$ 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 $\sqrt{n}\left(\psi_n^{TMLE} - \psi_0\right) \overset{D}{\to} \mathrm{N}(0,\sigma^2)$

AIPTW *a*ugmented *i*nverse *p*robability of *t*reatment *w*eighting

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

3 Longitudinal Data

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:

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

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

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

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

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

use SWIGs to visually check 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 $\Rightarrow$ $Y^g = Y \;\forall g \quad \Rightarrow \quad Y \perp\!\!\!\perp A_0 | L_0$ & $Y \perp\!\!\!\perp A_1 | A_0, L_0, L_1$; 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))

References

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

Hernán, M. A. and Robins, J. M. (2020). *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.