Causal Inference

a summary

Contents

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 most of 3

positivity: p. 155, p. 162

additional conditions: chapter 13.5

exchangeability: p 172f, p16-19

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

consistency: if $A = a$, then $Y^a = Y$ for each individual

$Y = Y^A$

technical point 3.2

effect modification chapter 4

interaction chapter 5

causal diagrams chapter 6, include swigs from 7.5 and that one technical point

more on SWIGS p 242ff

confounding chapter 7

selection bias chapter 8

measurement bias chapter 9

random variabilty chapter 10

2 Models

Modeling data are a sample from the target population

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

2.1 Traditional Methods

Outcome regression chapter 15

instrumental variable estimation chapter 16

causal survival analysis chapter 17 (and technical point 22.3)

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?

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

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

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

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

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

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

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

for *stochastic strategies* multiply with $\prod_{k=0}^{K} f^{int}\left(a_k \mid \bar{a}_{k-1}, \bar{l}_k\right)$
modelling:
iterated conditional expectation (ice)
$\mathrm{E}\left[Y_T^{\bar{a}}\right] = \mathrm{E}\left[\mathrm{E}\left[\mathrm{E}\left[...\mathrm{E}\left[Y_T \mid \bar{A}_{T-1}{=}\bar{a}_{T-1}, \bar{L}_T\right]...\mid \bar{A}_0{=}a_0, L_1\right] \mid L_0\right]\right]$
modelling:
g-null paradox

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

$$\begin{aligned}
\mathrm{E}\left[Y_1^{\bar{a}}\right] &\overset{\text{CE}}{=} \mathrm{E}\left[\mathrm{E}\left[Y_1^{\bar{a}} \mid A_0{=}a_0, L_0\right]\right] \\
\text{(ice)} &\overset{\text{CE*}}{=} \mathrm{E}\left[\mathrm{E}\left[\mathrm{E}\left[Y_1 \mid \bar{L}, \bar{A}{=}\bar{a}, Y_0\right] \mid A_0{=}a_0, L_0\right]\right] \\
&\overset{\text{LTP}}{=} \mathrm{E}\left[\textstyle\sum_{l_1} \mathrm{E}\left[Y_1 \mid A_0{=}a_0, \bar{L}, Y_0\right] \Pr\left[l_1 \mid a_0, l_0, y_0\right]\right] \\
&\overset{\text{LTP}}{=} \textstyle\sum_{l_0} \left[\sum_{l_1} \mathrm{E}\left[Y_1 \mid A_0{=}a_0, \bar{L}, Y_0\right] \Pr\left[l_1 \mid a_0, l_0, y_0\right]\right] \Pr\left[l_0\right] \\
\text{(jde)} &\overset{\text{sum}}{=} \textstyle\sum_{\bar{l}} \mathrm{E}\left[Y_1 \mid A_0{=}a_0, \bar{L}, Y_0\right] \Pr\left[l_1 \mid a_0, l_0\right] \Pr\left[l_0\right]
\end{aligned}$$

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

Marginal Structural Models

association is causation in the IP weighted pseudo-population

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

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. $\beta_0 + \beta_1 a + \beta_2 V a + \beta_3 V$; 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 $\mathrm{E}[Y|A=a] = \int \mathrm{E}[Y|L{=}l, A{=}a, C{=}0]\, dF_L[l]$

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

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

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

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

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)

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

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 ↓

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

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

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

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

find α_1 that is closest to zero

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:

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

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

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

$A_0 \to L_1 \to A_1$:

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))

target trial chapter 22 (does that even really fit in here, maybe push to 3rd paragraph in without models)

References

If no citation is given, the source is (?)

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.