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

Notation   chapter 1.1

average causal effect   chapter 1.2 and 1.3 and 1.4 and 1.5

randomized experiments (target trial)   2.1 and 2.2; 3.6

Standardization   plug-in (or parametric if so) g-formula
$$\mathrm{E}\left[Y|A=a\right] = \int \mathrm{E}\left[Y|L=l, A=a\right] dF_L\left[l\right]$$
weighted average of stratum-specific risks; unknowns can be
estimated non-parametrically or modeled

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

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

IP Weighting   adjust for (surrogate) confounders L
$$\mathrm{E}\left[Y|A=a\right] = \mathrm{E}\left[\frac{I(A=a)Y}{f\left[A|L\right]}\right]; W^A = \frac{1}{f\left[A|L\right]}; SW^A = \frac{f(A)}{f\left[A|L\right]}$$
unknowns can be estimated non-parametrically or modeled

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

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

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

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 $\downarrow$

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

identifiability conditions   most of 3
  positivity: p. 155, p. 162
  additional conditions: chapter 13.5
  exchangeability: p 172f
  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

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. $\mathrm{E}\left[Y|A=a\right]$
estimator:	function to use,	e.g. $\widehat{\mathrm{E}}\left[Y|A=a\right]$
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 $\rightarrow$ true value)

parsimonious model: few parameters estimate many quantities

bias-variance trade-off:
wiggliness $\uparrow$ $\rightarrow$ misspecification bias $\downarrow$, CI width $\uparrow$

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 $f\left[A\right]$) $\rightarrow$ 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

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
$\qquad\qquad SW^{A,C} = SW^A \times SW^C$ (uses $n^{c=0}$)

g-estimation can only adjust for confounding, not selection bias $\rightarrow$ use IP weights

G-Methods *g*eneralized treatment contrasts: adjust for (surrogate) confounders L

- **standardization**
- **IP weighting**
- **g-estimation:** not needed unless longitudinal

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

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

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

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 α_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

Outcome regression chapter 15

instrumental variable estimation chapter 16

causal survival analysis chapter 17

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

$\rightarrow$ **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?

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

    stochastic strategy: non-deterministic g

optimal strategy is where $\mathrm{E}[Y^g]$ 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 $\bar{a} = \bar{a}^*$; $Y^{\bar{a}} = Y$ if $\bar{A} = \bar{a}$;

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

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

use SWIGs to visually check d-separation

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

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 $\Rightarrow$

$Y^g = Y\ \forall g \quad \Rightarrow \quad Y \perp\!\!\!\perp A_0 | L_0$ & $Y \perp\!\!\!\perp A_1 | A_0, L_0, L_1$; therefore: if last two independences don't hold, one assumption is violated

g-null theorem: $\mathrm{E}[Y^g] = \mathrm{E}[Y]$, if the two independences hold ($\Rightarrow$ sharp null: only if strong faithfulness (no effect cancelling))

Standardization   over all possible $\bar{l}$-histories

simulates joint distribution of counterfactuals $\left(Y^{\bar{a}}, \bar{L}^{\bar{a}}\right)$ for $\bar{a}$

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

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

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

g-null paradox

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   reduce misspecification bias

    1. IP weights at each time m: $W^{\bar{A}_m} = \prod_{k=0}^{m} \frac{1}{f\left(A_k | \bar{A}_{k-1}, \bar{L}_k\right)}$

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

    3. calculate standardized mean outcome

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

g-estimation   chapter 21.4

censoring   chapter 21.5

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.