

On simulated and synthetic data for causal inference

STRATOS TG 7


Freiburg 2025











Bianca De Stavola, UCL (K); Els Goetghebeur, Gent (B);
Saskia Le Cessie, Leiden (NL); Erica Moodie, McGill (CA);
Ingeborg Waernbaum, Uppsala (S); Vanessa Didelez, Bremen (DE)


Building on Goetghebeur et al. 'Formulating causal questions and principled statistical answers' Stat Med, 2020.

github.com/IngWae/Formulating-causal-questions

The source code


IngWae
Add files via upload
eb758f0 · 5 years ago
5 Commits

 Analysis with SAS ...	Add files via upload	5 years ago
 Analysis with SAS ...	Add files via upload	5 years ago
 Analysis with SAS ...	Add files via upload	5 years ago
 Appendix2.R	Add files via upload	5 years ago
 Generate_simulati...	Add files via upload	5 years ago
 PROBITsim2018_v...	Add files via upload	5 years ago
 README.md	Create README.md	5 years ago
 stata_analyses_A...	Add files via upload	5 years ago
 stata_analyses_A...	Add files via upload	5 years ago
 stata_analyses_A...	Add files via upload	5 years ago


README

<https://github.com/IngWae/Formulating-causal-questions>

material for the paper Formulating causal questions and principled statistical answers

 Readme

 Activity

 2 stars

 1 watching

 0 forks

Report repository

Releases

No releases published

Packages

No packages published

Languages



03-06-2025, 09:06

GitHub - IngWae/Formulating-causal-questions: material for the paper Formulating causal questions and principled statistical answers

Formulating-causal-questions

This repository contains material for the paper: Formulating causal questions and principled statistical answers by Els Goetghebeur, Saskia le Cessie, Bianca de Stavola, Erica Moodie and Ingeborg Waernbaum

 R 38.8%
  Stata 31.5%
  SAS 29.7%

Why perform a simulation?

- To get a deeper understanding of data and methods to analyse the data.
- To show properties of a new method (e.g small sample behavior)
- To compare/optimize the performance of different methods under different conditions (cross validation is great for prediction (but not directly for causal prediction))
- To confirm calculations/analysis (i.e check (power) calculations, check an R-function)

All the more when you draw inference on *potential* outcomes

Morris et al. 'Using simulation studies to evaluate statistical methods', SIM 2019

The Simulation Learner

TG 7 wrote tutorial on causal questions and principled answers

- Overview of causal concepts and estimands
- with analysis methods to deal with timed exposures

Simulation Learner

- Simulated data inspired by existing trial
- Illustrates concepts and methods on counterfactual data

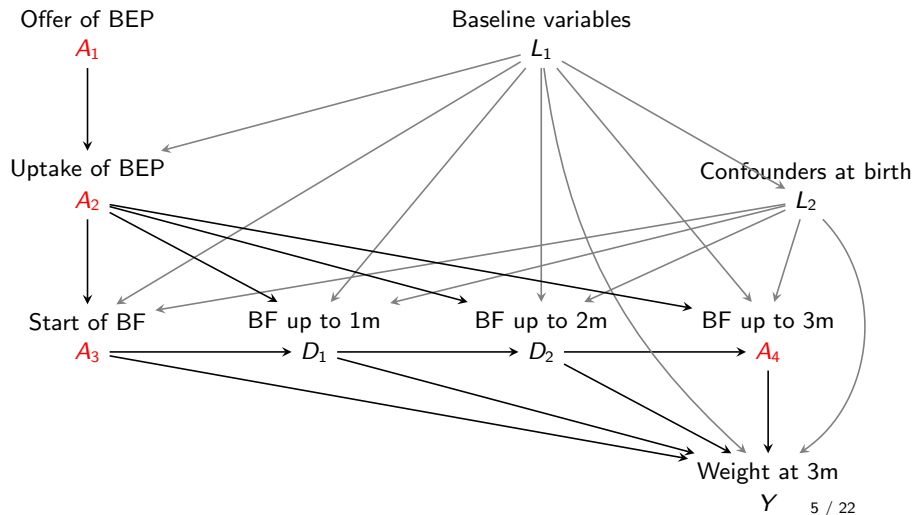
Promotion of Breastfeeding Intervention Trial - PROBIT. Kramer et al. (2001)

Promotion of Breastfeeding Intervention Trial - PROBIT

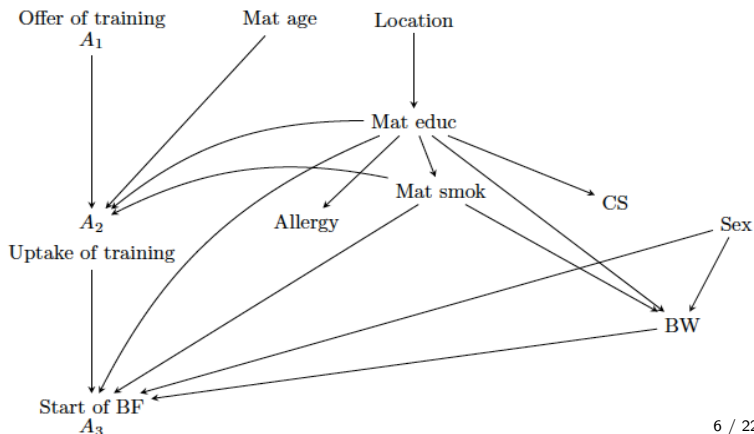
- Women living in a low income area of Belarus who gave birth to a full-term singleton baby between June '96- Dec '97
- were (cluster) **randomised to BF encouraging educational program** or not, during their last term of pregnancy.
- All babies were weighed at age 3 months.
- A simulated version of individually *randomised* women **Probitsim** included: 17,044 women with singleton births (8,667 in the active arm and 8,377 in the control arm).

Simulated on www.ofcaus.org — > [github](#)

Sketch of data generating model - causal DAG



Data generating model for 'observed' A_2 and A_3 given A_1 , L_1 , and L_2



The steps we take

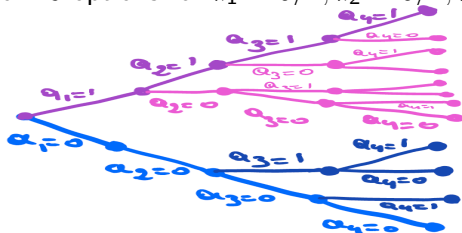
- Step 0: baseline covariate distribution L :
 - data as is (privacy breach?) or
 - simulated conditional discrete and continuous distributions
- Step 1: generate 'observed treatment'
 - Fit propensity model(s) $P(A_k = a | \mathcal{H}(\bar{A}_{k-}, \bar{Y}_{k-}))$
 - Simulate observed treatment(s) for every subject
- Step 2: generate (potential) outcomes :
 - fit outcome models given the history $\mathcal{H}(\bar{A}_k, \bar{Y}_{k-})$
 - for the observed treatment(s)
 - and for all *possible* alternative treatment histories.
- Step 3: sample out of these sequential (conditional) models
 - the set of counterfactual outcomes for each patient
 - e.g. 16 (12) sequences per patients in our case

The possible set of treatments

We simulate **the parallel worlds** for the PROBIT-like trial

- for the composition of the study population, i.e. mimic the baseline covariate distribution L in which we
- **'let' every person experience each of the possible exposure sequences :**

part of 16 options for $a_1 = 0/1, a_2 = 0/1, a_3 = 0/1, a_4 = 0/1$



- followed by the corresponding potential outcome(s)
- besides the potential exposures and outcomes, we **also simulate 'the observed' exposures and outcomes**

The steps we take

- Step 4: the result of 'setting treatment(s)' (a_k) according to that (possibly dynamic) rule ... $\rightarrow \bar{Y}_K(a - rule)$
 - simulate very large sample size once to get 'the population level value' \rightarrow gold standard
- Step 5: contrast mean outcomes under different treatments
 \Rightarrow estimates for the causal contrast
- Step 6: repeating for a 'dataset of estimated causal contrasts'
reveal (finite sample) bias and precision
- Step 7: generate *censoring* and other missing data patterns

Possible exposures a_1, a_2, a_3, a_4 relevant for whom?

When aiming to change

- the nature of the invitation to the breastfeeding program
- the content of the breastfeeding program to improve uptake of breastfeeding
- one's decision to start breastfeeding
- the supporting measures to improve maintaining breastfeeding for the full 3 months

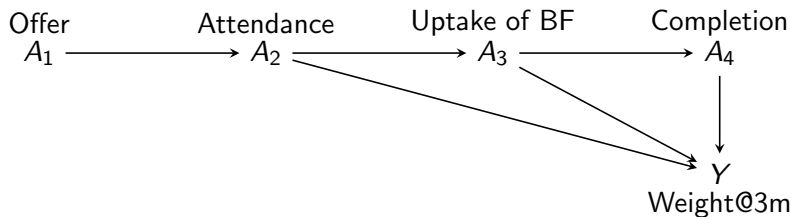
Special challenges

- ① Propagation of errors when extrapolating beyond the data (any treatment for all) – \rightarrow still plausible outcomes?
- ② Propose data generating models to check robustness
 - 1) consistent with our observations and plausible but
 - 2) substantially different so we can evaluate robustness of our methods to the pre-proposed methods
- ③ Missing data tied to intercurrent events (e.g. disease progression) with positivity problems
- ④ Inverse weighting/outcome regression/double robust methods/ Causal Machine learning/TMLE/...
- ⑤ 'Simple' analysis that can be based on summary statistics, like generalized linear models with or without Firth correction

Potential consequences of setting a_1

- $A_{2a_1(a)}$: the potential value of A_2 if A_1 is set to the value a .
- $A_{3a_1(a)} = 1$ would start BF if the programme were offered ($a = 1$) or not ($a = 0$).
- $A_{3a_1(a_1), a_2(a_2)} = 1$ would start BF if the programme were offered and followed ($a_1 = 1, a_2 = 1$), or offered but not followed ($a_1 = 1, a_2 = 0$) or not offered ($a_1 = 0, a_2 = 0$).

$$Y_{a_1(a)} = Y_{a_1(a)}(a_1 = a, A_{2a_1(a)}, A_{3a_1(a)}, A_{4a_1(a)})$$



Overall		
	$A_1 = 0$	$A_1 = 1$
	N (%)	N (%)
$A_2 = 0$	8377 (100)	3083 (35.6)
$A_2 = 1$	0 (0)	5584 (64.4)
$A_3 = 0$	4226 (50.5)	2782 (32.1)
$A_3 = 1$	4151 (49.5)	5885 (67.9)
All	8377 (100)	8667 (100)

Among those with $A_1 = 1$		
	$A_2 = 0$	$A_2 = 1$
	N (%)	N (%)
$A_3 = 0$	1745 (56.6)	1037 (18.6)
$A_3 = 1$	1338 (43.4)	4547 (81.4)
All	3083 (100)	5584 (100)

'Compliance' with program offer: cross world questions

Consider A_1 , the intervention of offering the program

2 types of compliers with the program offer:

- Program offer accepters: $\{A_{2a_1(1)} = 1 \text{ and } A_{2a_1(0)} = 0\}$
- BF compliers: $\{A_{3a_1(1)} = 1 \text{ and } A_{3a_1(0)} = 0\} = C$

Representing groups of women

- not directly observed
- most directly impacted by the intervention

'Feasible'? estimands for different purposes

The potential mean weight at three months in the study population under different possible conditions

outcome	interventions	pop	Education		
			low	int	high
$Y_{a_1(0)}$	BEP not offered	6017	5914	6057	6141
$Y_{a_1(1)}$	BEP offered	6115	6024	6155	6207
$Y_{a_2(1)}$	BEP followed	6182	6128	6208	6226
$Y_{a_3(0)}$	no BF	5827	5730	5854	5981
$Y_{a_1(0), a_3(1)}$	BEP not offered, BF started	6214	6154	6248	6246
$Y_{a_1(1), a_3(1)}$	BEP offered, BF started	6249	6207	6276	6262
$Y_{a_2(1), a_3(1)}$	BEP followed, BF started	6277	6261	6292	6266
$Y_{a_4(1)}$	duration BF = 3 months	6351	6393	6339	6286

The potential mean weight at three months in the study population under different possible conditions

Potential outcome	$A_2 = 1$	$A_1 = 1$ $A_2 = 0$	$A_1 = 1$ $A_3 = 1$	$A_1 = 1$ $A_3 = 0$	$A_1 = 0$ $A_3 = 1$	$A_1 = 0$ $A_3 = 0$
$Y_{a_1(0)}$	6047	5964	6149	5733	6274	5761
$Y_{a_1(1)}$	6200	5964	6292	5733	6308	5923
$Y_{a_2(1)}$	6200	6149	6308	5911	6329	6035
$Y_{a_3(0)}$	5849	5788	5871	5733	5893	5761
$Y_{a_1(0), a_3(1)}$	6226	6193	6251	6133	6274	6153
$Y_{a_1(1), a_3(1)}$	6282	6193	6292	6157	6308	6191
$Y_{a_2(1), a_3(1)}$	6282	6270	6308	6212	6329	6225
$Y_{a_4(1)}$	6345	6362	6372	6307	6392	6311

Estimated ATE and ATT of A_2 on weight at 3 months (in grams)

Estimand	Estimation method	Estimate	(SE)
ATE	True value	165.1	
	Crude regression	196.0	(9.6)
	Regression adjustment (without interactions)	155.4	(9.5)
	Regression adjustment (with interactions)	165.0	(9.7)
	PS stratification [†] (6 strata)	165.0	(9.4)
	Regression with PS [†]	156.2	(9.0)
	PS matching (1 match) [‡]	155.7	(10.1)
	PS matching (3 matches) [‡]	154.9	(10.1)
	PS IPW [†]	164.7	(9.3)
	PS DR IPW [†]	164.7	(9.7)
	IV	146.2	(14.0)
ATT	True value	152.8	
	Regression adjustment (with interactions)	148.7	(9.4)
	PS stratification [†] (6 strata)	148.7	(9.6)
	PS matching (1 match) [‡]	145.8	(9.8)
	PS matching (3 matches) [‡]	145.4	(9.7)
	PS IPW [†]	148.0	(9.6)

* controlled for: maternal age, maternal education, maternal allergy status, smoking status in the first trimester, and area of residence.

Results for A_3 (Starting breastfeeding)

Estimation method	$A_1 = 0$		$A_1 = 1$	
	Estimate	(SE)	Estimate	(SE)
ATE				
True value	386.8		422.3	
Crude regression	503.2	(11.6)	582.0	(12.2)
Regression (simple)	384.3	(2.8)	428.0	(3.3)
Regression (with interactions)	384.7	(3.2)	425.3	(2.7)
Regression with PS *	384.4	(3.2)	425.9	(3.3)
PS stratification* (6 strata)	392.2	(4.1)	442.0	(6.5)
PS matching (1 match)	386.5	(8.1)	429.0	(10.6)
PS matching (3 matches)	380.7	(5.5)	437.2	(7.8)
PS IPW	384.7	(3.8)	426.6	(6.7)
PS DR IPW	384.8	(3.9)	426.7	(7.0)
NO IV	513.3	(44.4)	—	—

Simulation learner is useful because:

- Generates observed data, augmented with potential outcomes
- Gives more insight in data generation process - *and* assumptions
- Actual causal effects are 'known' — > examine estimands
- Great help in finding correct ways of analysis (which turned out to be different for A_2 and A_3)
- Enables analytic methods comparisons (bias and precision)
- It is helpful in teaching causal methods
- Code of generation and analysis of data is available

Benchmarking Care Institutions

Let $Y(c)$ denote the 'outcome' that would have been observed for a given patient **if** treated at centre c .

Interest in the counterfactual risk $E\{Y(c)\} = E\{E\{Y(c)|L\}\}$ in centre c

- Regress outcome on patient characteristics affecting both outcome and center choice, then add a center effect
- this added effect is 'caused' by the center if all confounder are accounted for
- For quality of life outcome:
heavy undertaking, done some 10 years ago across Flanders.
- Individual centers intermediate benchmark – > summary stats

Other applications and simulations

TG7 and survival outcome - uses Rotterdam breast cancer data.
(Royston et al. BMC MRM, 2013)

- Violation of positivity assumption discovered
- Nice illustration of
 - missing confounders
 - impact of censoring not properly addressed

The SISAQOL project guidelines for QoL in late stage oncology
(Thomassen et al. BMC MRM, 24/25; Reynders et al. (in review))

- Repeated QoL and death as a 2-dimensional outcome
- Evaluate single arm trials with external control (SISAQOL) links with meta analysis and historical control simulation
- Contrast GEE and joint models for estimands and prediction

In 'conclusion'

- Between too simple and too complicated: chose wisely
eg. Niel Hens and infectious disease simulation in Flanders
(very detailed work for years)
- Bootstrapping and cross validation may be an option
- Causal questions open a whole new set of worlds
- We need more shared experience!
- TG7 repeated the tutorial - for survival outcomes, repeated
QoL and survival (including joint models)
- more to come on dynamic regimes

Thank you for a great workshop
and all the dedication to forward health and science!