Measuring the efficacy of the Counterweight Programme via g-computation algorithm

Marco Doretti, Sara Geneletti, Elena Stanghellini

London School of Economics, Department of Statistics
University of Perugia, Department of Economics

Abstract

- One of the purposes of longitudinal studies is the evaluation of the impact of a sequence of treatments/exposures on an outcome measured at the final stage;
- Time-varying confounding is one of the most important drawbacks that can arise when dealing with observational data: there may be variables that at each time act as both confounders and mediators;
- The g-computation algorithm is a popular method to overcome this issue;
- We propose an extension to deal with informative drop-out;
- The motivating example arises from the Counterweight Programme, a protocol developed in UK to tackle obesity.

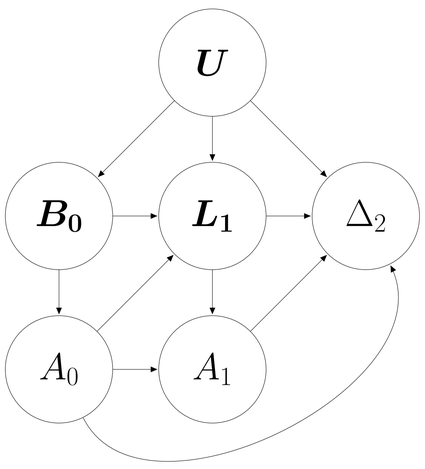
Framework

TREATMENT Focus is on *first line interventions*: in particular 'actions' (gym/diet prescription) are compared to 'discussion' (simple weight loss target setting);

CONFOUNDERS Time-invariant information about personal habits or clinical situation of patients (*baseline* variables) plus percentage changes in body mass index (*BMI*);

OUTCOME Percentage change in *BMI* at the final visit.

Causal graph and notation



Three temporal points are considered and the following notation is adopted:

- $\boldsymbol{B_0}$ is the set of baseline variables;
- (A_0, A_1) is the couple of treatment indicators;
- Δ_t is the percentage change in *BMI* at time t with respect to time 0;
- $\boldsymbol{L_1} = (\boldsymbol{B_0}, \Delta_1)$;
- $\boldsymbol{U}$ is the set of unobserved confounders;
- Δ_2 is the outcome.

Time-varying confounding

- The node $\boldsymbol{L_1}$ plays a *double role*: it is a *confounder* for the relation $A_1 - \Delta_2$ and a *mediator* for the relation $A_0 - \Delta_2$: a single regression model does not provide causal estimates for the effect of (A_0, A_1);
- The *no unmeasured confounding* assumption

$$\boldsymbol{U} \perp\!\!\!\perp A_0 \mid \boldsymbol{B_0}$$
$$\boldsymbol{U} \perp\!\!\!\perp A_1 \mid (A_0, \boldsymbol{L_1})$$

(*i.e.* no direct arrows from $\boldsymbol{U}$ to A_0 and A_1) is crucial for identifiability: it implies that only baseline variables and observed variations of *BMI* guide the decision of assigning/not assigning the treatment.

Algorithm settings

- The sequence of regressions was estimated on the complete cases ($n = 421$) relying on a *MAR* mechanism;
- Δ_t were simulated from $\mathcal{N}(\mu_t^s, \sigma_t^2)$ where the means μ_t^s depend on the particular strategy s under investigation and σ_t^2 are the residual variances from respective models;
- Implementation of stochastic dynamic regime shows a good overall fitting;
- All four treatments' combinations (static regime) and the deterministic dynamic regime (dd) 'treat until success' (decrease of at least 5% in *BMI* with respect to the initial value) were tested;
- For each strategy the estimated percentage change is reported together with bootstrap measures of uncertainty (standard error, 95% confidence interval and squared bias).

Problems

The *MAR* assumption is probably violated as subjects who leave the protocol early are likely to differ systematically from the others. The subsample of complete cases is self selected and therefore not representative, leading to poor estimates in the model for Δ_2.

Drop-out modeling

The formulation involving underlying continuous latent variables R_t^* is adopted: the two specifications are

$$\begin{cases} R_1^* = \boldsymbol{X_0}\boldsymbol{\alpha} + \eta_1^* \\ \Delta_1 = \boldsymbol{X_0}\boldsymbol{\beta} + \eta_1 \end{cases} \qquad \begin{cases} R_2^* = \boldsymbol{X_1}\boldsymbol{\gamma} + \eta_2^* \\ \Delta_2 = \boldsymbol{X_1}\boldsymbol{\delta} + \eta_2 \end{cases}$$

with $\boldsymbol{X_0} = (\boldsymbol{B_0}, A_0)$ and $\boldsymbol{X_1} = (\boldsymbol{B_0}, A_0, \Delta_1, A_1)$.
Selection equations are estimated through the observed indicators of attendance at visit $t = 1, 2$

$$R_t = \begin{cases} 1 \text{ if } R_t^* > 0 \\ 0 \text{ otherwise.} \end{cases}$$

As (Δ_1, A_1) are included as covariates in the second period equations, just *monotone drop-out patterns* can be considered.

Error terms

The following dependence structure for the error terms is assumed:

$$\begin{pmatrix} \eta_1^* \\ \eta_2^* \end{pmatrix} \sim \mathcal{N}_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \tilde{\rho} \\ \tilde{\rho} & 1 \end{pmatrix} \right)$$

and

$$\eta_1 = \sigma_{11}^* \eta_1^* + \epsilon_1 \qquad \sigma_{11}^* = Cov(\eta_1, \eta_1^*)$$
$$\eta_2 = \sigma_{22}^* \eta_2^* + \epsilon_2 \qquad \sigma_{22}^* = Cov(\eta_2, \eta_2^*)$$

where ϵ_1 and ϵ_2 are random deviations independent of all the other variables, without any further distributional assumption.

Adjustments

Letting $k_1 = \boldsymbol{X_0}\boldsymbol{\alpha}$ and $k_2 = \boldsymbol{X_1}\boldsymbol{\gamma}$, trivially

$$R_1 = 1 \Rightarrow R_1^* > 0 \Rightarrow \eta_1^* > -k_1$$
$$R_2 = 1 \Rightarrow R_2^* > 0 \Rightarrow \eta_2^* > -k_2.$$

so, under the assumptions made,

$$\mathbb{E}(\Delta_1 | \boldsymbol{X_0}, R_1 = 1) = \boldsymbol{X_0}\boldsymbol{\beta} + \sigma_{11}^* \lambda(k_1)$$

and

$$\mathbb{E}(\Delta_2 | \boldsymbol{X_1}, R_1 = 1, R_2 = 1) = \boldsymbol{X_1}\boldsymbol{\delta} + \sigma_{22}^* C_2(k_1, \tilde{\rho}, k_2)$$

where $\lambda(\cdot) = \frac{\phi(\cdot)}{\Phi(\cdot)}$ is the inverse Mills ratio (*i.e.* Heckman's two stage estimator) and

$$C_2(k_1, \tilde{\rho}, k_2) = \frac{\tilde{\rho}\phi(k_1)\left(1 - \Phi\left(\frac{\tilde{\rho}k_1 - k_2}{\sqrt{1-\tilde{\rho}^2}}\right)\right) + \phi(k_2)\left(1 - \Phi\left(\frac{\tilde{\rho}k_2 - k_1}{\sqrt{1-\tilde{\rho}^2}}\right)\right)}{\mathbb{P}(\eta_1^* > -k_1, \eta_2^* > -k_2)}$$

is the expectation of η_2^* in the bivariate truncated normal distribution.

Sensitivity analysis on $\tilde{\rho}$

- If $\tilde{\rho} = 0$ the drop-out processes at the two periods are independent;
- A cross validation approach provides acceptable estimates for $\tilde{\rho} \in [0.5; 0.8]$;
- Results are shown for $\tilde{\rho} = 0.5$.

Limitations

- *Exclusion restrictions* are necessary to avoid multicollinearity problems: information on age had to be omitted in the outcome equations;
- *Non-compliance* is likely to arise given treatment's definition.

Results: complete cases estimation

s	$\mathbb{E}(\Delta_2 \mid do(s))$	Std.Error	95% Conf.Interval		Bias^2
(0,0)	-4.68647	0.20616	-5.11907	-4.29964	0.00023
(0,1)	-4.47390	0.38022	-5.25351	-3.72289	0.00003
(1,0)	-5.51430	0.45071	-6.43527	-4.70392	0.00032
(1,1)	-5.31892	0.24370	-5.79154	-4.81332	0.00049
(dd)	-5.40825	0.27860	-5.98297	-4.87137	0.00001

Results: drop-out adjusted estimation

s	$\mathbb{E}(\Delta_2 \mid do(s))$	Std.Error	95% Conf.Interval		Bias^2
(0,0)	-4.18335	0.20516	-4.58789	-3.80654	0.00004
(0,1)	-4.31313	0.51957	-5.35391	-3.32939	0.00017
(1,0)	-4.95461	0.56221	-6.06241	-3.90563	0.00001
(1,1)	-5.08474	0.24577	-5.59197	-4.60780	0.00006
(dd)	-5.03836	0.30379	-5.67345	-4.47701	0.00008

Essential bibliography

[1] The Counterweight Project Team (2004) - *A new evidence-based model for weight management in primary care: the Counterweight Programme.*

[2] Robinson, J. M. (1986) - *A new approach to causal inference in mortality studies with a sustained exposure period-application to control of the healthy worker survivor effect.* Mathematical Modelling.

[3] Daniel, R. M. *et al.* (2012) - *Methods for dealing with time-dependent confounding.* Statistics in Medicine.