G-formula for Causal Inference via Multiple Imputation

Jonathan W. Bartlett

Department of Medical Statistics
LSHTM
www.thestatsgeek.com

ISCB 2023
Milan, Italy
28th August 2023

Acknowledgements

This is joint work with:

- Camila Olarte Parra, Emily Granger, Ruth Keogh (LSHTM)
- Erik van Zwet (Leiden)
- Rhian Daniel (Cardiff)

This work was supported by a UK Medical Research Council grant (MR/T023953/1).

G-formula

G-formula via multiple imputation

G-formula via MI with missing data

Conclusions

Outline

Time-varying treatments and confounders

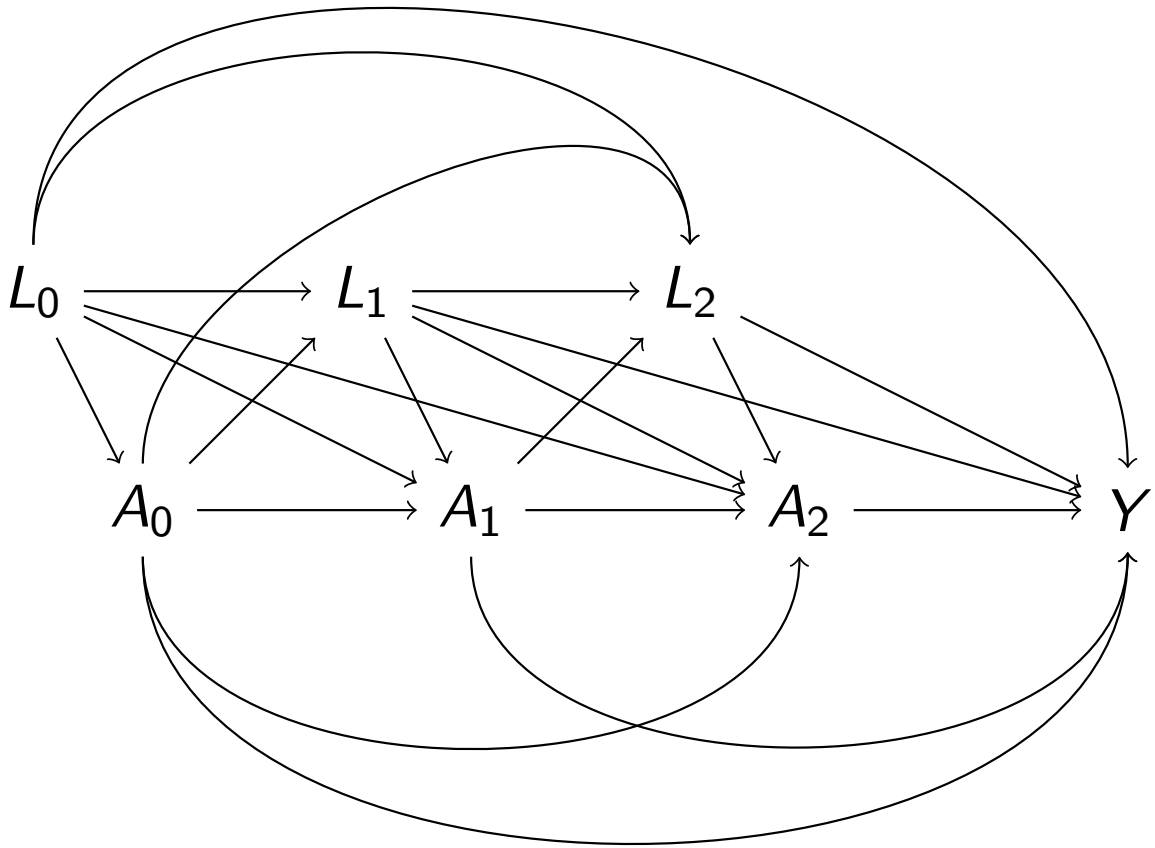
The setting under consideration is the 'standard' time-varying treatment and confounding setup.

A_k denotes treatment at time/visit k, $k = 0, \ldots, K$.

L_k denotes time-varying confounders at visit k.

Y denotes the final outcome of interest.

Directed acyclic graph (DAG)



G-formula

G-formula is one approach to estimation of quantities $E(Y^{\bar{a}}) = E(Y^{a_0,a_1,a_2})$ in this setting (see e.g. Ch. 21 of [1].

G-formula is based on the following equality (which follows from usual identification assumptions):

$$E(Y^{\bar{a}}) = \int_{l_0} \int_{l_1} \int_{l_2} E(Y|a_0, a_1, a_2, l_0, l_1, l_2) f(l_2|a_0, a_1, l_0, l_1) f(l_1|a_0, l_0) f(l_0) dl_2 dl_1 dl_0$$

This requires we specify and fit models for

- $f(L_0)$ (in fact, we typically empirically average across this, avoiding need for a model)
- $f(L_1|A_0, L_0)$
- $f(L_2|A_0, A_1, L_0, L_1)$
- $f(Y|A_0, A_1, A_2, L_0, L_1, L_2)$ (in fact, all we need is a model for $E(Y|A_0, A_1, A_2, L_0, L_1, L_2)$)

In general the integrals above are intractable.

Thus in practice implementations use Monte-Carlo integration.

G-formula by Monte-Carlo integration/simulation

To estimate $E(Y^{a_0,a_1,a_2})$, based on fitted models, for every individual we:

- simulate L_0^* from $f(L_0; \hat{\beta}_0)$ (or just use original, i.e. $L_0^* = L_0$)
- simulate L_1^* from $f(L_1|A_0 = a_0, L_0^*, \hat{\beta}_1)$
- simulate L_2^* from $f(L_2|A_0 = a_0, A_1 = a_1, L_0^*, L_1^*, \hat{\beta}_2)$
- simulate Y^* from $f(Y|A_0 = a_0, A_1 = a_1, A_2 = a_2, L_0^*, L_1^*, L_2^*, \hat{\beta}_Y)$ (or just calculate $E(Y|A_0 = a_0, A_1 = a_1, A_2 = a_2, L_0^*, L_1^*, L_2^*)$
- calculate mean of Y^* across individuals

  (or average $E(Y|A_0 = a_0, A_1 = a_1, A_2 = a_2, L_0^*, L_1^*, L_2^*)$ across individuals)

Outline

G-formula and imputation

From a missing data perspective, G-formula can be viewed as creating a single (improper) stochastic imputation of the longitudinal history under the treatment regime of interest.

In fact, to reduce Monte-Carlo error, implementations of G-formula create multiple imputations of these, and then average the imputed Y^* across individuals and across imputations.

For inference, implementations in Stata and R use non-parametric bootstrapping.

The close links begs the question - could we use existing (proper) multiple imputation software and Rubin's rules to perform G-formula?

G-formula via MI

For the longitudinal setup earlier, we can use MI to estimate $\mu = E(Y^{\bar{a}})$ in a G-formula type approach by:

1. Augment observed data with additional n_{syn} rows, setting L_0, L_1, L_2, Y to missing in the augmented rows to missing, and A_0, A_1, A_2 to value $\bar{a} = (a_0, a_1, a_2)$.
2. Run MI on the augmented dataset, generating M imputations.
3. For imputation m ($m = 1, \ldots, M$), calculate mean of Y from the augmented part of the dataset.
4. Average estimated means across M imputations (denoted $\hat{\mu}$) as estimator of $\mu = E(Y^{\bar{a}})$.

G-formula via MI - data structure

E.g. data structure for $\bar{a} = (1, 1, 1)$ is

R	L_0	A_0	L_1	A_1	L_2	A_2	Y
1	-0.3	0	0.5	0	2.2	1	1.3
1	2.3	1	4.2	1	4.6	1	5.5
1	-0.5	1	0.4	0	0.8	1	1.9
0	NA	1	NA	1	NA	1	NA
0	NA	1	NA	1	NA	1	NA
0	NA	1	NA	1	NA	1	NA

$R = 1$ indicates originally observed data

$R = 0$ indicates augmented data

G-formula via MI - implementation details

We have a block monotone missingness pattern in the augmented dataset.

Due to our earlier model assumptions, we can impute sequentially moving forwards in time:

1. Impute L^0
2. Impute $L_1|A_0, L_0$
3. Impute $L_2|A_0, A_1, L_0, L_1$
4. Impute $Y|A_0, A_1, A_2, L_0, L_1, L_2$

This means if we use for example chained equations MI software, there is no need to iterate around models.

We specify imputation equations as per above, and set iterations to 1.

Contrasts of treatment regimes

In practice we are interested in contrasts of the form $E(^{\bar{a}_1}) - E(^{\bar{a}_2})$ for regimes $\bar{a}_1$ and $\bar{a}_2$.

To estimate this, add augmented rows with $\bar{A} = \bar{a}_1$ and another set with $\bar{A} = \bar{a}_2$.

In the imputed datasets, calculate difference in sample means.

Inference for G-formula via MI estimator

How to estimate $\text{Var}(\hat{\mu})$ and conduct inference?

Ordinarily with MI we use Rubin's rules.

Estimate variance in each imputation and average these, yielding within-imputation variance $\hat{V}$.

Estimate variance of estimated means across M imputations, yielding between-imputation variance $\hat{B}$.

Then $\widehat{\text{Var}}(\hat{\mu}) = (1 + M^{-1})\hat{B} + \hat{V}$.

Unfortunately this does not work here - Rubin's variance estimator is much larger than the true $\text{Var}(\hat{\mu})$.

This is due to a form of uncongeniality - the imputation and analysis models are being fitted to different portions of the augmented dataset.

MI for synthetic samples/populations

Within the survey sampling field, there is an established literature on using MI to impute partially or fully synthetic datasets.

MI is used to impute/simulate variables for new/synthetic individuals, ensuring confidentiality of original participants.

In this context, Raghunathan, Reiter and Rubin [2] developed an alternative variance estimator:

$$(1 + M^{-1})\hat{B} - \hat{V}$$

where $\hat{B}$ and $\hat{V}$ are between and within-imputation variance.

In our pre-print, we use asymptotic theory for MI estimators of Robins and Wang [3] to show that Raghunathan *et al* 's variance estimator is asymptotically unbiased for the G-formula via MI estimator.

Outline

G-formula and missing data

In practice there will typically (always?!) be some missing data on baseline and time-varying confounders and treatment variables.

Given that MI can be used to perform G-formula when data are complete, can we use it to impute any missing data as well?

Yes. We impute the combination of the missing actual data and the missing potential outcome data (in the augmented part).

Imputing missing data with MI G-formula

There are (at least) two approaches:

1. **1.1** Augment observed data with extra rows as described earlier.
   1.2 Create M imputations of missing actual data and missing potential outcome data in one go.
2. **2.1** Impute missing actual data M times.
   2.2 Augment each imputed dataset with extra rows as described earlier.
   2.3 Impute missing potential outcomes in each augmented dataset once, giving M imputed datasets, and analyse as described earlier

We believe option 2 is more attractive, since we only have to impute the (usually) small amount of missing actual data to create a monotone pattern.

Outline

Conclusions

- MI seems like an attractive route to implementing G-formula, particularly when some data are missing.
- Simulation results so far suggest good performance of the synthetic combination rule.
- However, note $(1 + M^{-1})\hat{B} - \hat{V}$ can be negative - may need to increase n_{syn} or M.
- More details, simulation results, & data analysis, in our pre-print `https://arxiv.org/abs/2301.12026`
- R package gFormulaMI available on CRAN, which utilises `mice`.
- Here we considered static (fixed) treatment regimes. For dynamic treatment regimes, you can specify treatment rules via custom imputation method functionality in `mice` package.

References I

[1] M. Hernán and J. Robins.

Causal Inference: What If.

Boca Raton: Chapman & Hall/CRC, 2020.

[2] T. E. Raghunathan, J. P. Reiter, and D. B. Rubin.

Multiple imputation for statistical disclosure limitation.

Journal of Official Statistics, 19(1):1, 2003.

[3] J. M. Robins and N. Wang.

Inference for imputation estimators.

Biometrika, 85:113–124, 2000.