Multiple imputation for multilevel data with continuous and binary variables

V. Audigier[1,2,3], I. R. White[4,5], S. Jolani[6], T. P. A. Debray[7,8], M. Quartagno[9], J. Carpenter[9], S. van Buuren[10], M. Resche-Rigon[1,2,3]

1-Service de Biostatistique et Information Medicale, Hopital Saint-Louis, AP-HP, Paris, France 2-Universite Paris Diderot - Paris 7, Sorbonne Paris Cite, UMR-S 1153, Paris, France 3-INSERM, UMR 1153, Equipe ECSTRA, Hopital Saint-Louis, Paris 4-MRC Biostatistics Unit, Cambridge Institute of Public Health, U.K 5-MRC Clinical Trials Unit at UCL, London, U.K 6-Department of Methodology and Statistics, School of Public Health and Primary Care, Maastricht University, Maastricht, The Netherlands 7-Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands 8-Cochrane Netherlands, University Medical Center Utrecht, Utrecht, The Netherlands 9-Department of Medical Statistics, London School of Hygiene & Tropical Medicine, London, U.K. 10-Department of Statistics, TNO Prevention and Health, Leiden, The Netherlands

Abstract

Comparisons of multiple imputation (MI) methods for multilevel data with:

- binary and continuous variables
- systematically and sporadically missing values

Three dedicated methods are compared

- JM-jomo (Quartagno and Carpenter, 2016])
- FCS-GLM (Jolani et al., 2015)
- FCS-2stage (Resche-Rigon and White, 2016)

Performances of each MI method vary according to:

- the missing data pattern
- the multilevel structure
- the type of missing variables

Motivations: GREAT data

An IPD meta-analysis in cardiovascular disease consisting of 28 observational cohorts. Each study gathers a list of patient characteristics and potential risk factors (Lassus et al., 2013). → **Two-level data**

- 2 **binary** variables (AFIB and Gender)
- 7 **continuous** variables (BMI, Age, Systolic blood pressure (SBP), Diastolic blood pressure (DBP), Heart rate (HR), LVEF and BNP)
- 11685 individuals (between 18 and 1834 per study).
- each study is incomplete → **sporadically missing** values on all variables (except gender and LVEF)
- BNP, BMI, SBP, AFIB have not been collected for all studies → **systematically missing** values

Aim

Explaining the left ventricular ejection fraction (LVEF) from biomarkers that are easier to measure, such as BNP or electrocardiographic characteristics such as the atrial fibrillation (AFIB). For this, fit a generalized mixed model (GMM) → **estimate its fixed coefficients β and their associated variances $\widehat{\mathbf{Var}}\left[\hat{\beta}\right]$**

Multiple imputation methods

MI consists of three main steps:

1. imputing M times the data according to an imputation model
2. fitting a GMM to each imputed dataset $\rightarrow \left(\hat{\beta}_m, \widehat{\mathbf{Var}}\left[\hat{\beta}_m\right]\right)_{1\leq m\leq M}$
3. pooling estimates $\rightarrow \hat{\bar{\beta}}, \widehat{\mathbf{Var}}\left[\hat{\beta}\right]$

The imputation model needs to be in lines with the data:

- a GMM to account for the multilevel structure
- a link function to account for the type of variable
- identifiable parameters with systematically missing values

Imputation according to such a model requires reflecting the variability of its parameters $\boldsymbol{\theta}$. A Bayesian treatment of the model is used:

- prior distributions are assumed for $\boldsymbol{\theta}$
- based on the imputation model, posterior distributions are derived
- M draws of $\boldsymbol{\theta}$ are performed to impute the dataset M times

JM-jomo

- imputation model: a GMM model with multivariate outcome
- priors for $\boldsymbol{\theta}$: conjugate priors
- link function for binary variables: probit link
- implemented in the R package *jomo*

FCS-GLM

- imputation model: several GMM models with homoscedastic residuals (for continuous variables)
- priors for $\boldsymbol{\theta}$: Jeffrey priors
- link function for binary variables: logit link
- implemented in the R package *micemd*

FCS-2stage

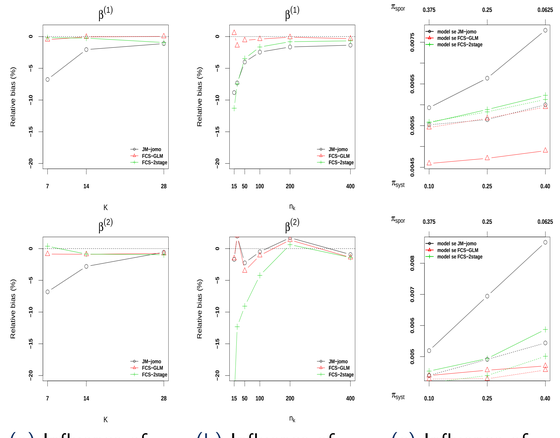
- imputation model: several GMM models with heteroscedastic residuals (for continuous variables)
- priors for $\boldsymbol{\theta}$: not defined, assumes large sample approximation for the posterior
- link function for binary variables: logit link
- implemented in the R package *micemd*

Experiments

500 multilevel incomplete datasets data are generated to mimic the GREAT data. Then, MI methods are applied on each one by using $M = 5$ imputed tables. Finally, the fixed coefficients of a GMM model $\beta = \left(\beta^{(1)}, \beta^{(2)}\right)$ and their associated variability are estimated

	Full	CC	FCS-noclust	FCS-fixclust	JM-jomo	FCS-GLM	FCS-2stage
$\beta^{(1)}$ est	-0.1101	-0.1104	-0.1039	-0.1102	-0.1088	-0.1100	-0.1090
$\beta^{(1)}$ rbias (%)	0.1	0.3	-5.6	0.2	-1.1	0.0	-0.9
$\beta^{(1)}$ model se	0.0047	0.0070	0.0041	0.0043	0.0066	0.0047	0.0059
$\beta^{(1)}$ emp se	0.0048	0.0071	0.0067	0.0058	0.0056	0.0057	0.0058
$\beta^{(1)}$ 95% cover	93.8	92.2	58.2	87.0	98.4	89.7	95.0
$\beta^{(1)}$ rmse	0.0048	0.0071	0.0091	0.0058	0.0058	0.0057	0.0059
$\beta^{(2)}$ est	0.0301	0.0299	0.0290	0.0300	0.0298	0.0298	0.0297
$\beta^{(2)}$ rbias (%)	0.2	-0.4	-3.4	0.0	-0.6	-0.8	-1.0
$\beta^{(2)}$ model se	0.0029	0.0053	0.0043	0.0043	0.0069	0.0046	0.0049
$\beta^{(2)}$ emp se	0.0030	0.0053	0.0045	0.0042	0.0049	0.0043	0.0044
$\beta^{(2)}$ 95% cover	94.2	94.4	92.0	94.6	97.6	95.8	96.2
$\beta^{(2)}$ rmse	0.0030	0.0053	0.0046	0.0042	0.0049	0.0043	0.0044
time (min)			0.9	1.1	7.8	103.3	0.9

Table: Inference results for the base-case configuration for: Full data (inference before deleting values), CC (complete-case analysis), FCS-noclust (ignoring the multilevel structure), FCS-fixclust (using fixed effect for the cluster), JM-jomo, FCS-GLM, FCS-2stage. Criteria related to the continuous (resp. binary) covariate are in light (resp. dark) grey. True values are $\beta^{(1)}$ =-0.11, $\beta^{(2)}$ =0.03.



(a) Influence of the number of clusters

(b) Influence of the cluster size

(c) Influence of the proportion of systematically missing values

Application

		CC	JM-jomo	FCS-GLM	FCS-2stage
β_{BNP}	est	-0.1132	-0.0891	-0.1002	-0.1009
	model se	0.0108	0.0078	0.0163	0.0112
β_{AFIB}	est	0.0268	0.0216	0.0218	0.0273
	model se	0.0071	0.0046	0.0066	0.0045
time (min)			94.0	30819.5	31.8

Table: Inference results for parameters of a GMM fitted on the GREAT data. 20 imputed arrays are considered for MI methods. 10 iterations are used for FCS methods. Estimates related to the continuous (resp. binary) covariate are in light (resp. dark) grey.

Discussion

- JM-jomo tends to overestimate the variance of the fixed coefficients because of unsuitable prior distributions
- FCS-2stage performs well with large clusters because of the large sample approximation
- FCS-GLM tends to underestimate the variance of the estimator because of the homoscedastic assumption for the residuals
- FCS methods provide larger variances than JM-jomo on the GREAT data. It could be due to convergence issues of FCS methods when the number of incomplete variables is larger

Guidelines

- include a large number of studies to obtain valid estimates
- use dedicated MI methods handling systematically missing values
- FCS-2stage performs well and quickly. It is notably relevant with a large proportion of systematically missing values, but should be avoided with small clusters.
- JM-jomo is recommended for large clusters when the proportion of binary variables is high
- FCS-GLM is recommended with small clusters

References

[1] V. Audigier, I. R. White, S. Jolani, T. P. A. Debray, M. Quartagno, J. Carpenter, S. van Buuren, and M. Resche-Rigon. Multiple imputation for multilevel data with continuous and binary variables. *ArXiv e-prints*, February 2017.

[2] Quartagno, M. and Carpenter, J. (2016). Multiple imputation for IPD meta-analysis: allowing for heterogeneity and studies with missing covariates. *Statistics in Medicine*, 35(17):2938–2954.

[3] Resche-Rigon, M. and White, I. (2016). Multiple imputation by chained equations for systematically and sporadically missing multilevel data. *Statistical Methods in Medical Research*.

[4] Jolani, S., Debray, T., Koffijberg, H., van Buuren, S., and Moons, K. (2015). Imputation of systematically missing predictors in an individual participant data meta-analysis: a generalized approach using MICE. *Statistics in Medicine*, 34(11):1841–1863.

[5] J. Lassus, E. Gayat, C. Mueller, W. Peacock, J. Spinar, VP. Harjola, R. van Kimmenade, A. Pathak, T. Mueller, and *et al.* Incremental value of biomarkers to clinical variables for mortality prediction in acutely decompensated heart failure: the MultinationalObservational Cohort on Acute Heart Failure (MOCA) study. *International Journal of Cardiology*, 168(3):2186–2194, October 2013.