

Multiple imputation for multilevel data with continuous and binary variables

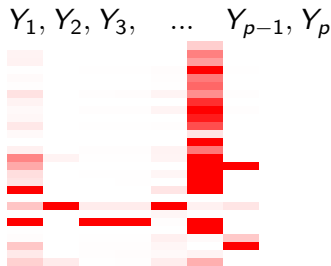
V. Audigier, I. White , S. Jolani, T. Debray, M. Quartagno, J. Carpenter, S. van Buuren, M. Resche-Rigon

CEDRIC, MSDMA team, CNAM, Paris

Chimiometrie XIX, 2018 January 31st, Paris

Motivation: GREAT data (GREAT Network, 2013)

- Risk factors associated with short-term mortality in acute heart failure
- 28 observational cohorts, 11685 patients, 2 **binary** and 8 **continuous** variables (patient characteristics and potential risk factors)
- sporadically and systematically **missing data**



Aim: explain the relationship between biomarkers (BNP, AFIB,...) and the left ventricular ejection fraction (LVEF)

$$y_{ik}^{LVEF} = \beta^0 + \beta^1 \mathbf{y}_{ik}^{BNP} + \beta^2 \mathbf{y}_{ik}^{AFIB} + b_k^0 + b_k^1 \mathbf{y}_{ik}^{BNP} + \varepsilon_{ik}$$

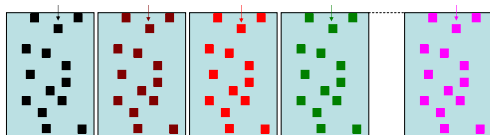
$$b_k \sim \mathcal{N}(0, \Psi) \quad \varepsilon_{ik} \sim \mathcal{N}(0, \sigma^2)$$

$\hat{\beta}$ and associated variability $\text{var}(\hat{\beta})$

Multiple imputation (Rubin, 1987)

- 1 Generate a set of M parameters $(\theta_m)_{1 \leq m \leq M}$ of an **imputation model** to generate M plausible imputed data sets

$$P(Y^{miss} | Y^{obs}, \theta_1) \quad \dots \quad P(Y^{miss} | Y^{obs}, \theta_M)$$



- 2 Fit the **analysis model** on each imputed data set: $\hat{\beta}_m, \widehat{\text{Var}}(\hat{\beta}_m)$

- 3 Combine the results: $\hat{\beta} = \frac{1}{M} \sum_{m=1}^M \hat{\beta}_m$

$$\widehat{\text{Var}}(\hat{\beta}) = \frac{1}{M} \sum_{m=1}^M \widehat{\text{Var}}(\hat{\beta}_m) + \left(1 + \frac{1}{M}\right) \frac{1}{M-1} \sum_{m=1}^M (\hat{\beta}_m - \hat{\beta})^2$$

⇒ **Provide estimation of the parameters and of their variability**

Imputation model for multilevel data

Two standard ways to perform MI

- **Fully conditional specification** (FCS, MICE): a conditional **imputation model** for each variable
- **Joint modelling** (JM): a joint **imputation model** for all variables

The **imputation model** (joint or conditional) needs to

- account for the **heterogeneity** between clusters
- account for the **types** of variables (continuous and binary)
- be identifiable with **sporadically and systematically** missing values

MI for multilevel data

Method (type - name)	deal with missing values:				Coded in R
	Spor.?	Syst.?	continuous ?	binary?	
JM-pan	yes	yes	yes	no	yes
JM-REALCOM	yes	yes	yes	yes	no
JM-jomo	yes	yes	yes	yes	yes
JM-Mplus	yes	yes	yes	yes	no
JM-RCME	yes	yes	yes	no	no
FCS-pan	yes	yes	yes	no	yes
FCS-2lnorm	yes	no	yes	no	yes
FCS-GLM	yes*	yes	yes	yes	yes
FCS-2stage	yes	yes	yes	yes*	yes

* using variant reported in this work

FCS-GLM (Jolani, 2017)

Conditional imputation models

$$y_{ik} = \mathbf{z}_{ik}\beta + \mathbf{w}_{ik}b_k + \varepsilon_{ik} \quad b_k \sim \mathcal{N}(0, \Psi) \quad \varepsilon_{ik} \sim \mathcal{N}(0, \sigma^2)$$

For each incomplete variable

- 1 generate $\theta_m = (\beta_m, \Psi_m, \sigma^2_m)$ $1 \leq m \leq M$
 - prior: non-informative (Jeffreys)
 - posterior distribution

$$\sigma^2 | Y, b \sim \text{Inv-}\Gamma \left(\frac{n-p}{2}, \frac{(n-p)\widehat{\sigma}^2}{2} \right)$$

$$\beta | Y, b, \sigma^2 \sim \mathcal{N}(\widehat{\beta}, \widehat{\text{var}}(\widehat{\beta}))$$

$$\Psi^{-1} | Y, b \sim \mathcal{W}(K, \widehat{bb}^\top)$$

- 2 impute in each cluster k with **systematically missing data**
 - draw $b_k \sim \mathcal{N}(0, \Psi_m)$
 - impute data according to the **imputation model**

FCS-GLM (Jolani, 2017)

Conditional imputation models

$$y_{ik} = \mathbf{z}_{ik}\beta + \mathbf{w}_{ik}b_k + \varepsilon_{ik} \quad b_k \sim \mathcal{N}(0, \Psi) \quad \varepsilon_{ik} \sim \mathcal{N}(0, \sigma^2)$$

For each incomplete variable

- 1 generate $\theta_m = (\beta_m, \Psi_m, \sigma^2_m)$ $1 \leq m \leq M$
 - prior: non-informative (Jeffreys)
 - posterior distribution

$$\sigma^2 | Y, b \sim \text{Inv-}\Gamma \left(\frac{n-p}{2}, \frac{(n-p)\widehat{\sigma}^2}{2} \right)$$

$$\beta | Y, b, \sigma^2 \sim \mathcal{N}(\widehat{\beta}, \widehat{\text{var}}(\widehat{\beta}))$$

$$\Psi^{-1} | Y, b \sim \mathcal{W}(K, \widehat{bb}^\top)$$

- 2 impute in each cluster k with **sporadically missing data**
 - draw $b_k \sim \mathcal{N}(\mu_{b_k|y_k}, \Psi_{b_k|y_k})$
 - impute data according to the **imputation model**

Differences between MI methods

	Prior	heteroscedasticity assumption	link binary
FCS-GLM	Jeffrey	no	logit
FCS-2stage		yes	logit
JM-jomo	conjugate	yes	probit

- conjugate prior distributions are known to very informative in GLMM
- heteroscedastic assumption is more flexible

Simulation design: data generation

500 incomplete data sets are independently simulated

- a multilevel structure
- 4 variables (1 binary, 3 continuous)
- sporadically and systematically missing data
- parameters are tuned to mimic GREAT data

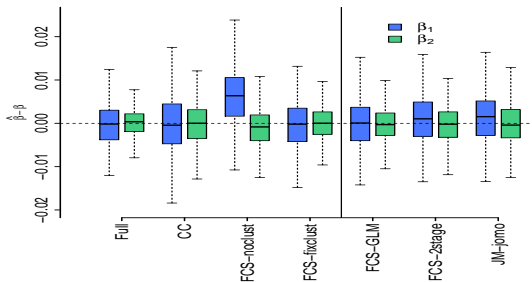
More precisely,

- $\mathbf{x}_{ik}^{(1)} : \mathcal{N}(2.9 + \mu_k, .36)$
- $\mathbf{x}_{ik}^{(2)} : \text{logit} \left(P \left(\mathbf{x}_{ik}^{(2)} = 1 \right) \right) = 4.2 + \nu_k \quad (\mu_k, \nu_k, \xi_k) \sim \mathcal{N} \left(0, \begin{bmatrix} .12 & .001 & .001 \\ .001 & .12 & .001 \\ .001 & .001 & .12 \end{bmatrix} \right)$
- $\mathbf{x}_{ik}^{(3)} : \mathcal{N}(2.9 + \xi_k, .36)$
- $y_{ik} = \beta^0 + \beta^1 \mathbf{x}_{ik}^{(1)} + \beta^2 \mathbf{x}_{ik}^{(2)} + b_k^0 + b_k^1 \mathbf{x}_{ik}^{(1)} + \varepsilon_{ik}$
with $\beta = (.72, -.11, .03)$, $\Psi = \begin{bmatrix} .0077 & .0015 \\ .0015 & .0004 \end{bmatrix}$, $\sigma = .15$
- add missing values on $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}$ with $\pi_{\text{sys}} = .25$ and $\pi_{\text{spor}} = .25$

Simulation design

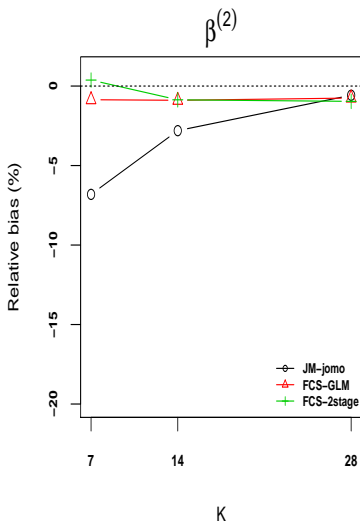
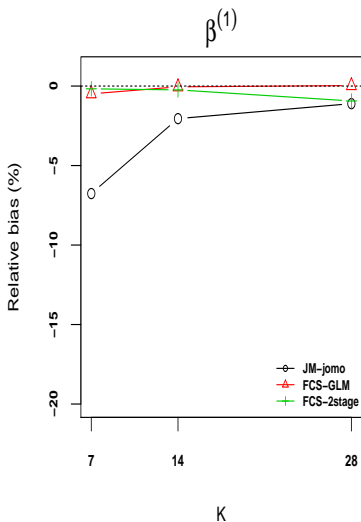
- **Methods**
 - JM-jomo, FCS-GLM, FCS-2stage
 - Full, CC, FCS-fixclust, FCS-noclust
 - $M = 5$ imputed arrays
- **Estimands:** β and $var(\hat{\beta})$
- **Criteria:** bias, rmse, variance estimate, coverage

Results: base-case

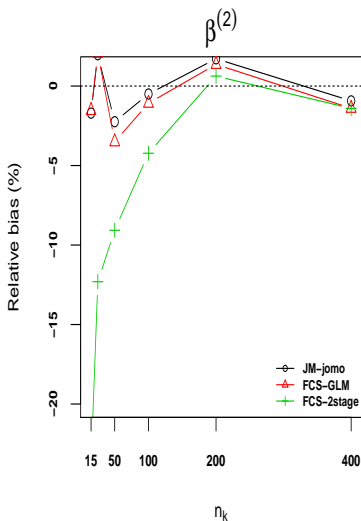
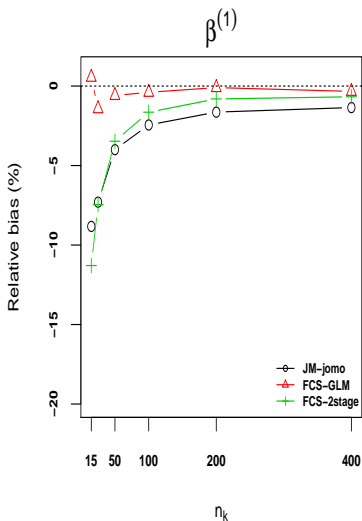


Method	$\sqrt{\widehat{\text{var}}(\hat{\beta})}$		$\sqrt{\text{var}(\hat{\beta})}$		95% Cover		Time (min)
	β_1	β_2	β_1	β_2	β_1	β_2	
Full	0.0047	0.0029	0.0048	0.0030	93.8	94.2	
CC	0.0070	0.0053	0.0071	0.0053	92.2	94.4	
FCS-noclust	0.0041	0.0043	0.0067	0.0045	58.2	92.0	0.9
FCS-fixclust	0.0043	0.0043	0.0058	0.0042	87.0	94.6	1.1
FCS-GLM	0.0047	0.0046	0.0057	0.0043	89.7	95.8	103.3
FCS-2stage	0.0059	0.0049	0.0058	0.0044	95.0	96.2	0.9
JM-jomo	0.0066	0.0069	0.0056	0.0049	98.4	97.6	7.8

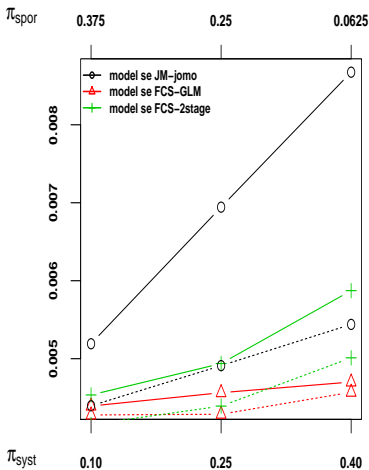
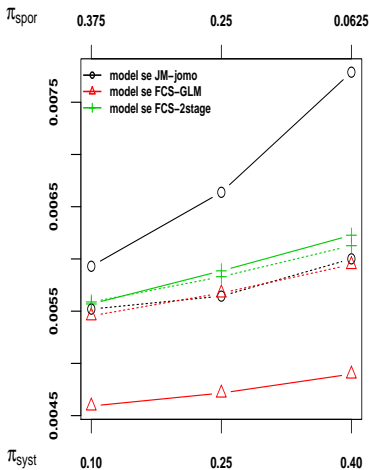
Influence of the number of clusters



Influence of the cluster size



Influence of the proportion of systematically missing values



Conclusion

An overview of MI methods for multilevel data

- FCS-GLM, FCS-2stage and JM-jomo all appear to perform well
- Outperform had-hoc methods

FCS-2stage

- provides a quick way to obtain first results
- for large clusters

FCS-GLM

- tends to underestimate the variance of the estimator because of the homoscedastic assumption
- recommended with small clusters
- time consuming (with binary variables)

JM-jomo

- tends to overestimate the variance of the fixed coefficients because of unsuitable prior distributions
- recommended for large clusters when the proportion of binary variables is high

Methods are implemented in R packages (mice, micemd, jomo)

References I

- Global Research on Acute conditions Team (GREAT) Network. Managing Acute Heart Failure in the ED - Case Studies from the Acute Heart Failure Academy, 2013. <http://www.greatnetwork.org>.
- D. Rubin. *Multiple Imputation for Non-Response in Survey*. Wiley, New-York, 1987.
- S. Jolani. Hierarchical imputation of systematically and sporadically missing data: An approximate bayesian approach using chained equations. *Biometrical Journal*, 2017. doi: 10.1002/bimj.201600220.
- 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*, 2017.
- V. Audigier and M. Resche-Rigon. *micemd: Multiple Imputation by Chained Equations with Multilevel Data*, 2017. R package version 1.2.0.
- M. Quartagno and J. Carpenter. *jomo: A package for Multilevel Joint Modelling Multiple Imputation*, 2016. URL <http://CRAN.R-project.org/package=jomo>. R package version 2.2-0.
- Matthieu Resche-Rigon and Ian White. Multiple imputation by chained equations for systematically and sporadically missing multilevel data. *Statistical Methods in Medical Research*, 2016. <http://dx.doi.org/10.1177/0962280216666564>.