micemd: a smart multiple imputation R package for missing multilevel data

V. Audigier, M. Resche-Rigon
CEDRIC, MSDMA team, CNAM, Paris

UseR, 2019 July 11th, Toulouse
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)

\[
y_{ik}^{LVEF} = \beta^0 + \beta^1 y_{ik}^{BNP} + \beta^2 y_{ik}^{AFIB} + b_k^0 + b_k^1 y_{ik}^{BNP} + \varepsilon_{ik}
\]

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

\[\hat{\beta} \text{ and associated variability } \text{var}\left(\hat{\beta}\right)\]
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

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

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

3. Combine the results:

$$\hat{\beta} = \frac{1}{M} \sum_{m=1}^{M} \hat{\beta}_m$$

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

$\Rightarrow$ 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

<table>
<thead>
<tr>
<th>Method</th>
<th>Spor.?</th>
<th>Syst.?</th>
<th>continuous?</th>
<th>binary?</th>
<th>Coded in R</th>
</tr>
</thead>
<tbody>
<tr>
<td>JM-pan</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>JM-REALCOM</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>JM-jomo</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>JM-Mplus</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>JM-RCME</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>FCS-pan</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>FCS-2lnorm</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>FCS-GLM</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>FCS-2stage</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
</tbody>
</table>
FCS-2stage (Resche-Rigon and White, 2016)

Conditional imputation models

\[ y_{ik} = z_{ik} (\beta + b_k) + \varepsilon_{ik} \]
\[ b_k \sim \mathcal{N}(0, \Psi) \quad \varepsilon_{ik} \sim \mathcal{N}(0, \sigma_k^2) \]

1. generate \( \theta_m = (\beta_m, \Psi_m, (\sigma_1^2, \ldots, \sigma_K^2)_m) \)

   - estimate \( \theta \) and \( \text{var} (\hat{\theta}) \) with a two-stage estimator

   stage 1 fit \( y_{ik} = z_{ik} \beta_k + \varepsilon_{ik} \) to each cluster

   stage 2 combine the \( \hat{\beta}_k \) (and \( \hat{\sigma}_k^2 \)) with a random intercept model:

   \[ \hat{\beta}_k = (\beta + b_k) + \varepsilon'_k \quad b_k \sim \mathcal{N}(0, \Psi) \quad \varepsilon'_k \sim \mathcal{N}(0, \text{var}(\hat{\beta}_k)) \]

   - draw \( \theta_m \) from the asymptotic distribution of the estimator with expectation \( \hat{\theta} \) and variance \( \hat{\text{var}}(\hat{\theta}) \)

2. impute in each cluster \( k \) with systematically missing data

   - draw \( b_k \) from their marginal distribution
   - impute data according to the imputation model
FCS-2stage (Resche-Rigon and White, 2016)

**Conditional imputation models**

\[ y_{ik} = z_{ik} (\beta + b_k) + \varepsilon_{ik} \quad b_k \sim \mathcal{N}(0, \Psi) \quad \varepsilon_{ik} \sim \mathcal{N}(0, \sigma_k^2) \]

1. generate \( \theta_m = (\beta_m, \Psi_m, (\sigma_1^2, \ldots, \sigma_K^2)_m) \)
   - estimate \( \theta \) and \( \text{var} (\hat{\theta}) \) with a **two-stage estimator**

   **stage 1** fit \( y_{ik} = z_{ik} \beta_k + \varepsilon_{ik} \) to each cluster
   **stage 2** combine the \( \hat{\beta}_k \) (and \( \hat{\sigma}_k^2 \)) with a random intercept model:

\[ \hat{\beta}_k = (\beta + b_k) + \varepsilon'_k \quad b_k \sim \mathcal{N}(0, \Psi) \quad \varepsilon'_k \sim \mathcal{N}(0, \text{var} (\hat{\beta}_k)) \]

   - draw \( \theta_m \) from the **asymptotic** distribution of the estimator with expectation \( \hat{\theta} \) and variance \( \hat{\text{var}} (\hat{\theta}) \)

2. impute in each cluster \( k \) with **sporadically missing data**
   - draw \( b_k \) **conditionally to** \( \hat{\beta}_k \) from stage 2
   - impute data according to the imputation model
Differences between MI methods (Audigier et al., 2018)

<table>
<thead>
<tr>
<th></th>
<th>Prior</th>
<th>heteroscedasticity assumption</th>
<th>link</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCS-2stage</td>
<td>yes</td>
<td></td>
<td>logit</td>
</tr>
<tr>
<td>FCS-GLM</td>
<td>Jeffrey</td>
<td>no</td>
<td>logit</td>
</tr>
<tr>
<td>JM-jomo</td>
<td>conjugate</td>
<td>yes</td>
<td>probit</td>
</tr>
</tbody>
</table>

- 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,

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

- **Methods**
  - JM-jomo, FCS-GLM, FCS-2stage
  - Full, CC, FCS-fixclust, FCS-noclust
  - $M = 5$ imputed arrays

- **Estimands**: $\beta$ and $\text{var}\left(\hat{\beta}\right)$

- **Criteria**: bias, rmse, variance estimate, coverage
Results: base-case

<table>
<thead>
<tr>
<th>Method</th>
<th>$\sqrt{\text{var} \left( \hat{\beta} \right)}$ $\beta_1$</th>
<th>$\sqrt{\text{var} \left( \hat{\beta} \right)}$ $\beta_2$</th>
<th>$\beta_1$ 95% Cover</th>
<th>$\beta_2$ 95% Cover</th>
<th>Time (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full</td>
<td>0.0047</td>
<td>0.0029</td>
<td>93.8</td>
<td>94.2</td>
<td></td>
</tr>
<tr>
<td>CC</td>
<td>0.0070</td>
<td>0.0053</td>
<td>92.2</td>
<td>94.4</td>
<td></td>
</tr>
<tr>
<td>FCS-noclust</td>
<td>0.0041</td>
<td>0.0043</td>
<td>58.2</td>
<td>92.0</td>
<td>0.9</td>
</tr>
<tr>
<td>FCS-fixclust</td>
<td>0.0043</td>
<td>0.0043</td>
<td>87.0</td>
<td>94.6</td>
<td>1.1</td>
</tr>
<tr>
<td>FCS-GLM</td>
<td>0.0047</td>
<td>0.0046</td>
<td>89.7</td>
<td>95.8</td>
<td>103.3</td>
</tr>
<tr>
<td>FCS-2stage</td>
<td>0.0059</td>
<td>0.0049</td>
<td>95.0</td>
<td>96.2</td>
<td>0.9</td>
</tr>
<tr>
<td>JM-jomo</td>
<td>0.0066</td>
<td>0.0069</td>
<td>98.4</td>
<td>97.6</td>
<td>7.8</td>
</tr>
</tbody>
</table>
Influence of the number of clusters

\begin{align*}
\beta^{(1)}(K) & \quad \text{Relative bias (\%)} \\
7 & \quad -20 \\
14 & \quad -15 \\
28 & \quad 0
\end{align*}

\begin{align*}
\beta^{(2)}(K) & \quad \text{Relative bias (\%)} \\
7 & \quad -20 \\
14 & \quad -15 \\
28 & \quad 0
\end{align*}

JM-jomo, FCS-GLM, FCS-2stage
Influence of the cluster size

\( \beta^{(1)} \)

\( \beta^{(2)} \)

Relative bias (%)

\( \eta_k \)

JM-jomo
FCS-GLM
FCS-2stage

\( \eta_k \)
Influence of the proportion of systematically missing values

\[ \pi_{\text{syst}} \]

\[ \pi_{\text{spor}} \]

\[ 0.0045 \quad 0.0055 \quad 0.0065 \quad 0.0075 \]

\[ 0.10 \quad 0.25 \quad 0.40 \]

\[ 0.0045 \quad 0.005 \quad 0.006 \quad 0.007 \quad 0.008 \]

\[ 0.10 \quad 0.25 \quad 0.40 \]

- model se JM-jomo
- model se FCS-GLM
- model se FCS-2stage
micemd is an add-on for the mice R package \(^1\) which performs multiple imputation by chained equations.

**Imputation models** for
- multilevel data
- sporadically and systematically missing values
- continuous, binary or count variables

**Tools** to facilitate its use
- automatic choice of imputation models
- choice of the number of imputed tables
- overimputation for model checking
- parallel calculation

\(^1\)van Buuren and Groothuis-Oudshoorn (2011)
References


