Multiple imputation for multilevel data with continuous and binary variables


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)

\[
y_{ik}^{LVEF} = \beta_0 + \beta_1 y_{ik}^{BNP} + \beta_2 y_{ik}^{AFIB} + b_k^0 + b_k^1 y_{ik}^{BNP} + \epsilon_{ik}
\]

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

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

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

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

3. Combine the results:

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

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

⇒ Provide estimation of the parameters and of their variability
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>

* using variant reported in this work
Conditional imputation models

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

For each incomplete variable

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

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

\[
\beta | Y, b, \sigma^2 \sim \mathcal{N} \left( \hat{\beta}, \text{var} \left( \hat{\beta} \right) \right)
\]

\[
\Psi^{-1} | Y, b \sim \mathcal{W} \left( K, \hat{bb}^\top \right)
\]

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

$$ y_{ik} = z_{ik}\beta + w_{ik}b_k + \varepsilon_{ik} \quad b_k \sim \mathcal{N} (0, \Psi) \quad \varepsilon_{ik} \sim \mathcal{N} (0, \sigma^2) $$

For each incomplete variable

1. generate $\theta_m = (\beta_m, \Psi_m, \sigma^2_m)$ \hspace{2cm} $1 \leq m \leq M$
   - prior: non-informative (Jeffreys)
   - posterior distribution
   $$
   \sigma^2 | Y, b \sim \text{Inv-\Gamma} \left( \frac{n - p}{2}, \frac{(n - p)\hat{\sigma}^2}{2} \right)
   $$
   $$
   \beta | Y, b, \sigma^2 \sim \mathcal{N} \left( \hat{\beta}, \text{var} (\hat{\beta}) \right)
   $$
   $$
   \Psi^{-1} | Y, b \sim \mathcal{W} \left( K, \hat{b} \hat{b}^\top \right)
   $$

2. impute in each cluster $k$ with sporadically missing data
   - draw $b_k \sim \mathcal{N} (\mu_{b_k|y_k}, \Sigma_{b_k|y_k})$
   - impute data according to the imputation model
## Differences between MI methods

<table>
<thead>
<tr>
<th></th>
<th>Prior</th>
<th>Heteroscedasticity Assumption</th>
<th>Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCS-GLM</td>
<td>Jeffrey</td>
<td>No</td>
<td>Logit</td>
</tr>
<tr>
<td>FCS-2stage</td>
<td>Yes</td>
<td>Logit</td>
<td></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 be 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,

- \( x_{ik}^{(1)} \sim \mathcal{N}(2.9 + \mu_k, .36) \)
- \( x_{ik}^{(2)} : \frac{1}{\pi_{syst}} \logit \left( P \left( x_{ik}^{(2)} = 1 \right) \right) = 4.2 + \nu_k \) \( (\mu_k, \nu_k, \xi_k) \sim \mathcal{N} \left( 0, \begin{bmatrix} .12 & .001 & .001 \\ .001 & .12 & .001 \\ .001 & .001 & .12 \end{bmatrix} \right) \)
- \( x_{ik}^{(3)} \sim \mathcal{N}(2.9 + \xi_k, .36) \)
- \( y_{ik} = \beta^0 + \beta^1 x_{ik}^{(1)} + \beta^2 x_{ik}^{(2)} + b_k^0 + b_k^1 x_{ik}^{(1)} + \varepsilon_{ik} \) 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 \( var(\hat{\beta}) \)

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

<table>
<thead>
<tr>
<th>Method</th>
<th>$\sqrt{\text{var} (\hat{\beta})}$</th>
<th>$\sqrt{\text{var} (\hat{\beta})}$</th>
<th>$95%$ Cover</th>
<th>Time (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\beta_1$</td>
<td>$\beta_2$</td>
<td>$\beta_1$</td>
<td>$\beta_2$</td>
</tr>
<tr>
<td>Full</td>
<td>0.0047</td>
<td>0.0029</td>
<td>0.0048</td>
<td>0.0030</td>
</tr>
<tr>
<td>CC</td>
<td>0.0070</td>
<td>0.0053</td>
<td>0.0071</td>
<td>0.0053</td>
</tr>
<tr>
<td>FCS-noclust</td>
<td>0.0041</td>
<td>0.0043</td>
<td>0.0067</td>
<td>0.0045</td>
</tr>
<tr>
<td>FCS-fixclust</td>
<td>0.0043</td>
<td>0.0043</td>
<td>0.0058</td>
<td>0.0042</td>
</tr>
<tr>
<td>FCS-GLM</td>
<td>0.0047</td>
<td>0.0046</td>
<td>0.0057</td>
<td>0.0043</td>
</tr>
<tr>
<td>FCS-2stage</td>
<td>0.0059</td>
<td>0.0049</td>
<td>0.0058</td>
<td>0.0044</td>
</tr>
<tr>
<td>JM-jomo</td>
<td>0.0066</td>
<td>0.0069</td>
<td>0.0056</td>
<td>0.0049</td>
</tr>
</tbody>
</table>
Influence of the number of clusters

- $\beta^{(1)}$ vs. $K$
- $\beta^{(2)}$ vs. $K$

Relative bias (%) vs. $K$ for different methods:
- JM-jomo
- FCS-GLM
- FCS-2stage
Influence of the cluster size

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

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

Relative bias (%)

- JM-jomo
- FCS-GLM
- FCS-2stage

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

\[ \pi_{\text{syst}} \quad 0.10 \quad 0.25 \quad 0.40 \]
\[ \pi_{\text{spor}} \quad 0.375 \quad 0.25 \quad 0.0625 \]

- model se JM-jomo
- model se FCS-GLM
- model se FCS-2stage

Comparisons
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)


