EP16: Missing Values in Clinical Research: Multiple Imputation

10. Requirements for MICE to work (well)

Nicole Erler

Department of Biostatistics, Erasmus Medical Center

n.erler@erasmusmc.nl

Erasmus MC
University Medical Center Rotterdam

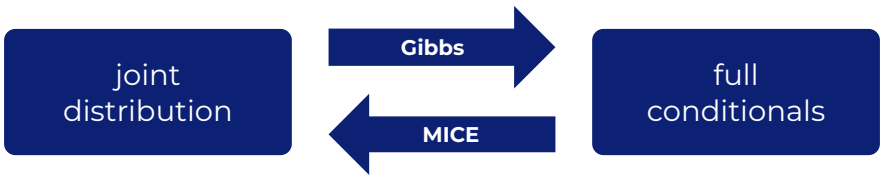
Joint and Conditional Distributions

Recall: The MICE algorithm is based on the idea of Gibbs sampling.

Joint and Conditional Distributions

Recall: The MICE algorithm is based on the idea of Gibbs sampling.

Gibbs sampling exploits the fact that a joint distribution is fully determined by its full conditional distributions.



In MICE, the full conditionals are not derived from the joint distribution: we directly specify the full conditionals and hope a joint distribution exists.

Joint and Conditional Distributions

The **uncertainty about whether a joint distribution exists** for the specified set of imputation models is often considered to be mainly a theoretical problem.

In practice, violations only have little impact on results in many applications.

Joint and Conditional Distributions

The **uncertainty about whether a joint distribution exists** for the specified set of imputation models is often considered to be mainly a theoretical problem.

In practice, violations only have little impact on results in many applications.

However, as we have seen in the examples on the previous slides, there are **settings where the direct specification** of the full conditionals/imputation models **may lead to problems**, causing biased results.

Some Conditions and Definitions

Two important definitions:

Compatibility:

A joint distribution exists, that has the full conditionals (imputation models) as its conditional distributions.

Congeniality:

The imputation model is compatible with the analysis model.

Some Conditions and Definitions

Important requirements for MICE to work well include:

- Compatibility

Some Conditions and Definitions

Important requirements for MICE to work well include:

- Compatibility
- Congeniality

Some Conditions and Definitions

Important requirements for MICE to work well include:

- Compatibility
- Congeniality
- MAR or MCAR (in the standard implementations)

Some Conditions and Definitions

Important requirements for MICE to work well include:

- Compatibility
- Congeniality
- MAR or MCAR (in the standard implementations)
- **All relevant variables** need to be included. (Omission might result in MNAR.)

Some Conditions and Definitions

Important requirements for MICE to work well include:

- Compatibility
- Congeniality
- MAR or MCAR (in the standard implementations)
- **All relevant variables** need to be included. (Omission might result in MNAR.)
- **The outcome needs to be included** as predictor variable (but we usually do not impute missing outcome values).

Some Conditions and Definitions

Important requirements for MICE to work well include:

- Compatibility
- Congeniality
- MAR or MCAR (in the standard implementations)
- **All relevant variables** need to be included. (Omission might result in MNAR.)
- **The outcome needs to be included** as predictor variable (but we usually do not impute missing outcome values).
- The imputation models (and analysis model) need to be **correctly specified** (which is a requirement in any standard analysis).

Why imputation with MICE can go wrong

What went wrong in our previous examples?

Why imputation with MICE can go wrong

What went wrong in our previous examples?

When incomplete variables have **non-linear associations** with the outcome, or with each other, the requirement(s) of ***compatibility*** **and/or** ***congeniality*** **are violated**.

Why imputation with MICE can go wrong

What went wrong in our previous examples?

When incomplete variables have **non-linear associations** with the outcome, or with each other, the requirement(s) of ***compatibility*** **and/or** ***congeniality*** **are violated**.

Omission, or inadequate inclusion, of the outcome may result in **MNAR** missing mechanisms. The same is the case when other relevant predictor variables are not used as predictor variables in the imputation.

Why imputation with MICE can go wrong

What went wrong in our previous examples?

When incomplete variables have **non-linear associations** with the outcome, or with each other, the requirement(s) of ***compatibility*** **and/or** ***congeniality*** **are violated**.

Omission, or inadequate inclusion, of the outcome may result in **MNAR** missing mechanisms. The same is the case when other relevant predictor variables are not used as predictor variables in the imputation.

Furthermore, **omission of variables** may lead to **mis-specified models**, however, models may also be mis-specified when all relevant covariates are included, but **distributional assumptions** or the specified **form of associations** are incorrect.

Alternatives to MICE

To **avoid incompatible** and **uncongenial** imputation models, we need to

- specify the joint distribution
- and derive full conditionals / imputation models from this joint distribution

instead of specifying them directly.

Alternatives to MICE

To **avoid incompatible** and **uncongenial** imputation models, we need to

- specify the joint distribution
- and derive full conditionals / imputation models from this joint distribution

instead of specifying them directly.

Problem:
The joint distribution may not be of any known form:

$$\begin{array}{l} x_1 \sim N(\mu_1, \sigma_1^2) \\ x_2 \sim N(\mu_2, \sigma_2^2) \end{array} \quad \Rightarrow \quad \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \right)$$

but $$\begin{array}{l} x_1 \sim N(\mu_1, \sigma_1^2) \\ x_2 \sim Bin(\mu_2) \end{array} \quad \Rightarrow \quad \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim ???$$

Alternatives to MICE

Possible approaches:

Approach 1: **Multivariate Normal Model**
Approximate the joint distribution by a known multivariate distribution.

(usually the normal distribution; this is the approach mentioned in Section 01)

Approach 2: **Sequential Factorization**
Factorize the joint distribution into a (sequence of) conditional and a marginal distributions.

Multivariate Normal Model

Assumption:
The outcome and incomplete variables follow a **joint multivariate normal distribution**, conditional on the completely observed covariates $\mathbf{X}_c$, parameters $\boldsymbol{\theta}$ and, possibly, random effects, $\mathbf{b}$:

$$p(\mathbf{y}, \mathbf{x}_1, \ldots, \mathbf{x}_p \mid \mathbf{X}_c, \boldsymbol{\theta}, \mathbf{b}) \sim N(\boldsymbol{\mu}, \Sigma)$$

Multivariate Normal Model

Assumption:
The outcome and incomplete variables follow a **joint multivariate normal distribution**, conditional on the completely observed covariates $\mathbf{X}_c$, parameters θ and, possibly, random effects, $\mathbf{b}$:

$$p(\mathbf{y}, \mathbf{x}_1, \ldots, \mathbf{x}_p \mid \mathbf{X}_c, \theta, \mathbf{b}) \sim N(\mu, \Sigma)$$

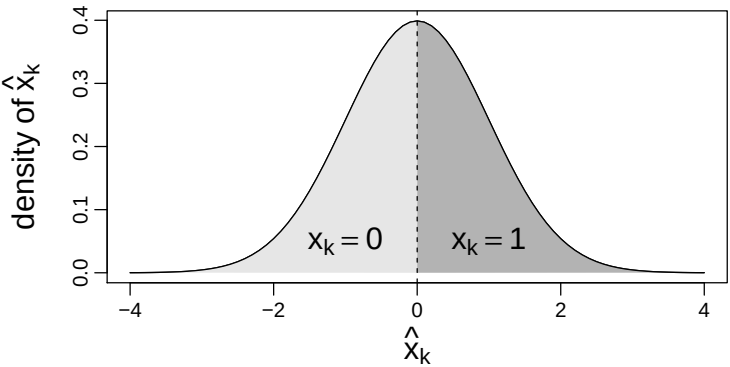
How do we get that multivariate normal distribution?

1. Assume **all** incomplete variables and the outcome are **(latent) normal**.
2. Specify linear (mixed) **models based on observed covariates**.
3. **Connect** using multivariate normal for **random effects & error terms**.

Multivariate Normal Model

1. Latent normal assumption:

e.g.: $\mathbf{x}_k$ binary $\rightarrow$ latent $\hat{\mathbf{x}}_{\boldsymbol{k}}$ is standard normal: $\begin{cases} \mathbf{x}_k = 1 & \text{if } \hat{\mathbf{x}}_{\boldsymbol{k}} \geq 0 \\ \mathbf{x}_k = 0 & \text{if } \hat{\mathbf{x}}_{\boldsymbol{k}} < 0 \end{cases}$



Multivariate Normal Model

2. Specify models:

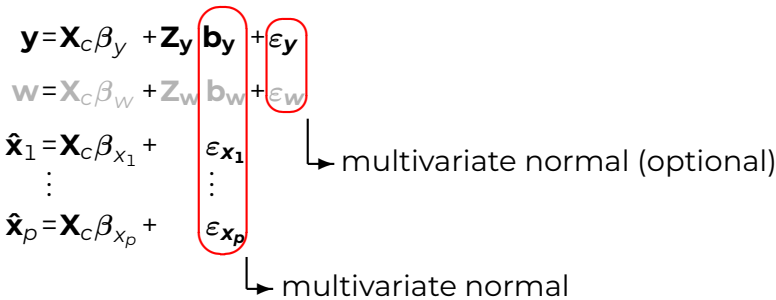
$$\mathbf{y} = \mathbf{X}_c\beta_y + \mathbf{Z_y}\,\mathbf{b_y} + \varepsilon_{\boldsymbol{y}}$$

$$\mathbf{w} = \mathbf{X}_c\beta_w + \mathbf{Z_w}\,\mathbf{b_w} + \varepsilon_{\boldsymbol{w}}$$

$$\begin{aligned} \hat{\mathbf{x}}_1 &= \mathbf{X}_c\beta_{x_1} + \varepsilon_{\boldsymbol{x_1}} \\ &\vdots \qquad\qquad \vdots \\ \hat{\mathbf{x}}_p &= \mathbf{X}_c\beta_{x_p} + \varepsilon_{\boldsymbol{x_p}} \end{aligned}$$

Multivariate Normal Model

2. Specify models / 3. Connect random effects & error terms:



Multivariate Normal Model

Advantages:

- easy to specify
- relatively easy to implement
- relatively easy to sample from
- works for longitudinal outcomes

Disadvantages:

- assumes linear associations

Imputation with **non-linear associations** or **survival data** is possible with **extensions** of the multivariate normal approach.

Sequential Factorization

The **joint distribution** of two variables y and x can be written as the product of conditional distributions:

$$p(y, x) = p(y \mid x)\, p(x)$$

(or alternatively $p(y, x) = p(x \mid y)\, p(y)$)

Sequential Factorization

The **joint distribution** of two variables y and x can be written as the product of conditional distributions:

$$p(y, x) = p(y \mid x)\, p(x)$$

(or alternatively $p(y, x) = p(x \mid y)\, p(y)$)

This can easily be **extended for more variables**:

$$p(y, x_1, \ldots, x_p, X_c) = \underbrace{p(y \mid x_1, \ldots, x_p, X_c)}_{\text{analysis model}}\, p(x_1 \mid x_2, \ldots, x_p, X_c)\ \ldots\ p(x_p \mid X_c)$$

where $x_1, \ldots, x_p$ denote incomplete covariates and X_c contains all completely observed covariates.

Sequential Factorization

The **analysis model** is part of the specification of the joint distribution.

- ➡ The **outcome**
  - is **automatically included in the imputation** procedure
  - does not appear in any of the predictors of the imputation models:
  - ➡ **no need to approximate/summarize** complex outcomes!

Sequential Factorization

The **analysis model** is part of the specification of the joint distribution.

- ➡ The **outcome**
  - is **automatically included in the imputation** procedure
  - does not appear in any of the predictors of the imputation models:

  ➡ **no need to approximate/summarize** complex outcomes!

- ➡ The **parameters of interest** are obtained directly

  ➡ Imputation and analysis in one step!

Sequential Factorization

The **analysis model** is part of the specification of the joint distribution.

- ➡ The **outcome**
  - is **automatically included in the imputation** procedure
  - does not appear in any of the predictors of the imputation models:

  ➡ **no need to approximate/summarize** complex outcomes!

- ➡ The **parameters of interest** are obtained directly

  ➡ Imputation and analysis in one step!

- ➡ **Non-linear associations / interactions**
  - specified in the analysis model

  ➡ **automatically taken into account!**

Sequential Factorization

The **analysis model** is part of the specification of the joint distribution.

- ➡ The **outcome**
  - is **automatically included in the imputation** procedure
  - does not appear in any of the predictors of the imputation models:

  ➡ **no need to approximate/summarize** complex outcomes!

- ➡ The **parameters of interest** are obtained directly

  ➡ Imputation and analysis in one step!

- ➡ **Non-linear associations / interactions**
  - specified in the analysis model

  ➡ **automatically taken into account!**

Since the joint distribution usually does not have a known form, Gibbs sampling is used to estimate parameters and sample imputed values.

Sequential Factorization

Advantages:

- **flexible:**
  - any outcome type
  - separate imputation models per variable
- can handle **non-linear associations** and interactions
- assures **congeniality and compatibility**

Disadvantages:

- specification takes requires time and consideration
- sampling may be more **computationally intensive**

Some Relevant R Packages

For complex settings there are alternatives to **mice**:

For example the R packages **JointAI**, **smcfcs** and **jomo**.

Some Relevant R Packages

For complex settings there are alternatives to **mice**:

For example the R packages **JointAI**, **smcfcs** and **jomo**.

- they use **Bayesian methodology** to impute values

Some Relevant R Packages

For complex settings there are alternatives to **mice**:

For example the R packages **JointAI**, **smcfcs** and **jomo**.

- they use **Bayesian methodology** to impute values
- **jomo** and **smcfcs** perform **multiple imputation**;
  the imputed datasets that can then be analysed the same way data imputed by **mice** would be analysed.

Some Relevant R Packages

For complex settings there are alternatives to **mice**:

For example the R packages **JointAI**, **smcfcs** and **jomo**.

- they use **Bayesian methodology** to impute values
- **jomo** and **smcfcs** perform **multiple imputation**;
  the imputed datasets that can then be analysed the same way data imputed by **mice** would be analysed.
- **JointAI** works **fully Bayesian**
  - performs analysis and imputation simultaneously
  - ➡ results from the analysis model of interest are obtained directly

R package smcfcs

Substantive Model Compatible Fully Conditional Specification, a hybrid approach between FCS and sequential factorization (Bartlett et al. 2015)

smcfcs (version 1.5.0) can impute incomplete covariates in

- linear regression
- logistic regression
- poisson regression
- Weibull survival models
- Cox proportional hazard models
- competing risk survival models
- nested case control studies
- case cohort studies

while ensuring compatibility between analysis model and imputation models.

For more information see the help files and the vignette.

R Package jomo

JOint MOdel imputation using the multivariate normal approach, with **extensions to assure compatibility** between analysis and imputation models. (Carpenter and Kenward 2012)

jomo (version 2.7-2) can handle

- linear regression
- generalized linear regression
- proportional odds (ordinal) probit regression
- linear mixed models
- generalized linear mixed models
- (ordinal) cumulative link mixed models
- Cox proportional hazards models.

For more info see the help file.

R Package JointAI

Joint Analysis and Imputation,
uses the **sequential factorization approach** to perform simultaneous analysis and imputation. (Erler et al. 2016, 2019)

JointAI (version 1.0.2) can analyse incomplete data using

- linear regression
- generalized linear regression
- linear mixed models
- generalized linear mixed models
- (ordinal) cumulative logit regression
- (ordinal) cumulative logit mixed models
- parametric (Weibull) survival models
- Cox proportional hazards models

while assuring compatibility between analysis model and imputation models when non-linear functions or interactions are included.

R Package JointAI

The necessary **Gibbs sampling** is performed using **JAGS** (an external program), which is free, but needs to be installed from https://sourceforge.net/projects/mcmc-jags/files/.

JointAI can be installed from CRAN or GitHub (development version containing bug fixes and other improvements)

```
install.packages("devtools")
devtools::install_github("NErler/JointAI")
```

JointAI has its own web page (https://nerler.github.io/JointAI/) with several vignettes on Visualization of Incomplete Data, a Minimal Example, details on Model Specification, etc.

References I

Bartlett, Jonathan W, Shaun R Seaman, Ian R White, James R Carpenter, and Alzheimer's Disease Neuroimaging Initiative. 2015. "Multiple Imputation of Covariates by Fully Conditional Specification: Accommodating the Substantive Model." *Statistical Methods in Medical Research* 24 (4): 462–87. https://doi.org/10.1177/0962280214521348.

Carpenter, James, and Michael Kenward. 2012. *Multiple Imputation and Its Application*. John Wiley & Sons. https://doi.org/10.1002/9781119942283.

Erler, Nicole S, Dimitris Rizopoulos, Vincent WV Jaddoe, Oscar H Franco, and Emmanuel MEH Lesaffre. 2019. "Bayesian Imputation of Time-Varying Covariates in Linear Mixed Models." *Statistical Methods in Medical Research* 28 (2): 555–68. https://doi.org/10.1177/0962280217730851.

References II

Erler, Nicole S, Dimitris Rizopoulos, Joost van Rosmalen, Vincent WV Jaddoe, Oscar H Franco, and Emmanuel MEH Lesaffre. 2016. “Dealing with Missing Covariates in Epidemiologic Studies: A Comparison Between Multiple Imputation and a Full Bayesian Approach.” *Statistics in Medicine* 35 (17): 2955–74. https://doi.org/10.1002/sim.6944.