EP16: Missing Values in Clinical Research: Multiple Imputation

4. A Closer Look at the Imputation Step

Nicole Erler

Department of Biostatistics, Erasmus Medical Center

n.erler@erasmusmc.nl

Erasmus MC
University Medical Center Rotterdam

The Imputation Step

The imputation step consists itself of two (or three) steps:

0. specification of the imputation model
1. **estimation** / sampling **of the parameters**
2. **drawing imputed values** from the predictive distribution

The Imputation Step

The imputation step consists itself of two (or three) steps:

0. specification of the imputation model
1. **estimation** / sampling **of the parameters**
2. **drawing imputed values** from the predictive distribution

Notation:

y: variable to be imputed

X: design matrix of other variables

$$\mathbf{y} = \begin{matrix} \mathbf{y}_{obs} \left\{ \vphantom{\begin{matrix} y_1 \\ \vdots \\ y_q \end{matrix}} \right. \\ \mathbf{y}_{mis} \left\{ \vphantom{\begin{matrix} NA \\ \vdots \\ NA \end{matrix}} \right. \end{matrix} \begin{bmatrix} y_1 \\ \vdots \\ y_q \\ NA \\ \vdots \\ NA \end{bmatrix}$$

$$\mathbf{X} = \begin{matrix} \mathbf{X}_{obs} \left\{ \vphantom{\begin{matrix} x \\ \vdots \\ x \end{matrix}} \right. \\ \mathbf{X}_{mis} \left\{ \vphantom{\begin{matrix} x \\ \vdots \\ x \end{matrix}} \right. \end{matrix} \begin{bmatrix} x_{11} & \dots & x_{1p} \\ \vdots & \dots & \vdots \\ x_{q1} & \dots & x_{qp} \\ x_{q+1,1} & \dots & x_{q+1,p} \\ \vdots & \dots & \vdots \\ x_{n1} & \dots & x_{np} \end{bmatrix}$$

Bayesian Multiple Imputation

In the **Bayesian framework:**
everything unknown or unobserved is considered a **random variable**.

For example:

- regression coefficients β,
- residual variance σ^2 and
- missing values $\mathbf{y}_{mis}$.

Bayesian Multiple Imputation

In the **Bayesian framework:**
everything unknown or unobserved is considered a **random variable**.

For example:

- regression coefficients β,
- residual variance σ^2 and
- missing values $\mathbf{y}_{mis}$.

Random variables have a **probability distribution**.

- The **expectation** of that distribution quantifies which **values** of the random variable are **most likely**.
- The **variance** is a measure of the **uncertainty** about the values.

Bayesian Multiple Imputation

In **Bayesian imputation**:

1. in the **observed data**:
   estimate the distribution of **the parameters** describing the association between incomplete variables and the other data

$$p(\mathbf{y}_{obs} \mid \mathbf{X}_{obs}, \beta, \sigma) \quad \Rightarrow \quad p(\beta \mid \mathbf{y}_{obs}, \mathbf{X}_{obs}),\ p(\sigma \mid \mathbf{y}_{obs}, \mathbf{X}_{obs})$$

2. use these estimates to obtain the the probability **distribution of incomplete variables** given the other data

$$p(\mathbf{y}_{mis} \mid \mathbf{X}_{mis}, \beta, \sigma)$$

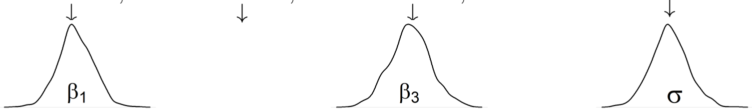
3. **sample values** from these distributions ➡ **imputation**

Bayesian Multiple Imputation

Step 1:

Specify a (Bayesian) regression model

$$y_{obs} = \beta_0 + \beta_1 x_{1,obs} + \beta_2 x_{2,obs} + \beta_3 x_{3,obs} + \ldots + \varepsilon$$

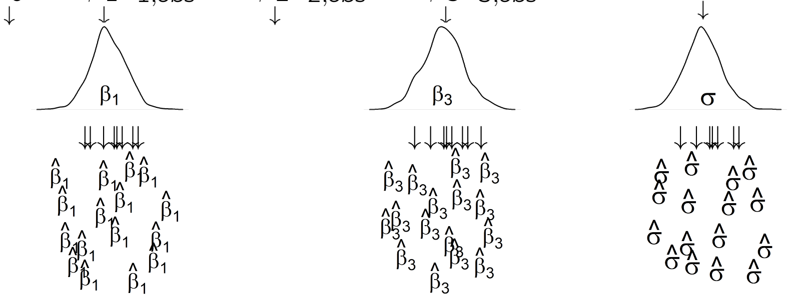


Bayesian Multiple Imputation

Step 1:

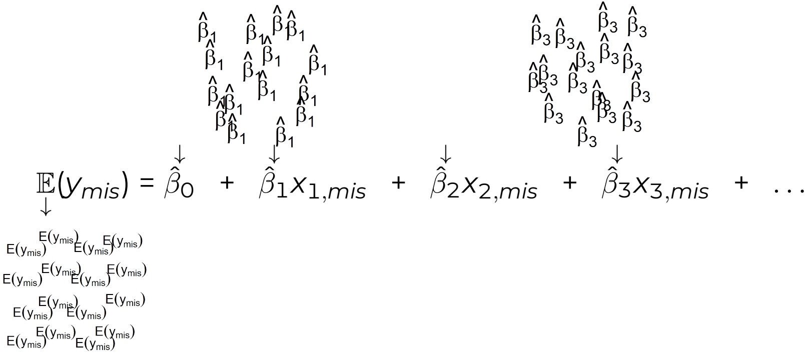
Specify a (Bayesian) regression model

$$y_{obs} = \beta_0 + \beta_1 x_{1,obs} + \beta_2 x_{2,obs} + \beta_3 x_{3,obs} + \ldots + \varepsilon$$



Bayesian Multiple Imputation

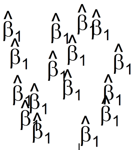
Step 2:

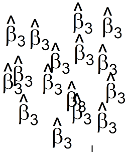


$$\mathbb{E}(y_{mis}) = \hat{\beta}_0 + \hat{\beta}_1 x_{1,mis} + \hat{\beta}_2 x_{2,mis} + \hat{\beta}_3 x_{3,mis} + \ldots$$

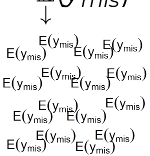
Bayesian Multiple Imputation

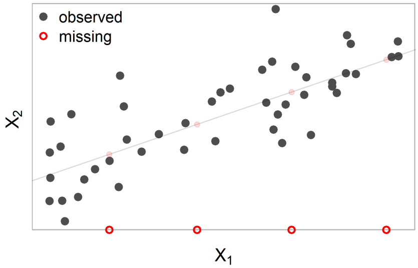
Step 2:





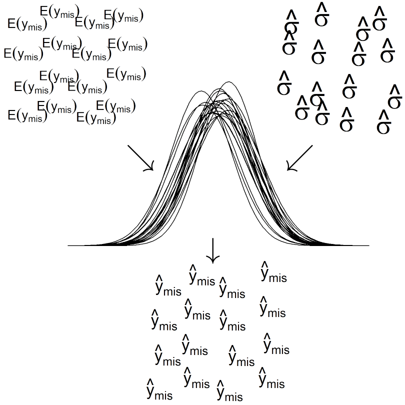
$$\mathbb{E}(y_{mis}) = \hat{\beta}_0 + \hat{\beta}_1 x_{1,mis} + \hat{\beta}_2 x_{2,mis} + \hat{\beta}_3 x_{3,mis} + \dots$$

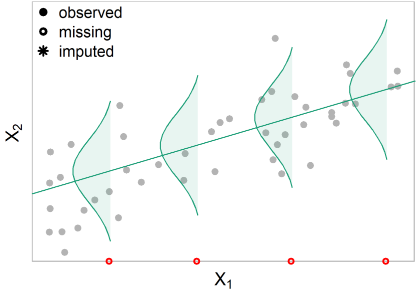




Bayesian Multiple Imputation

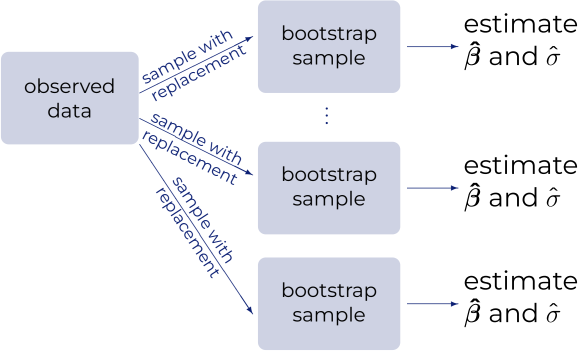
Step 3:





Bootstrap Multiple Imputation

Alternative approach to capture the uncertainty: **bootstrap**



Bootstrap samples can contain some **observations multiple times** and some **observations not at all**.

Bootstrap Multiple Imputation

In **bootstrap multiple imputation**,

- per imputation: **one bootstrap sample** of the **observed data**
- the (least squares or maximum likelihood) estimates of the parameters are calculated from

$$\mathbf{y}_{obs} = \underset{\substack{\downarrow \\ \hat{\beta}}}{\mathbf{X}_{obs}\beta} + \underset{\substack{\downarrow \\ \hat{\sigma}}}{\varepsilon_{obs}} \qquad \text{(step 1).}$$

- Imputed values are sampled from $p(\mathbf{y}_{mis} \mid \mathbf{X}_{mis}, \hat{\beta}, \hat{\sigma})$ (step 2).

Bootstrap Multiple Imputation

In **bootstrap multiple imputation**,

- per imputation: **one bootstrap sample** of the **observed data**
- the (least squares or maximum likelihood) estimates of the parameters are calculated from

$$\underset{}{\mathbf{y}_{obs}} = \mathbf{X}_{obs}\underset{\hat{\beta}}{\underset{\downarrow}{\beta}} + \underset{\hat{\sigma}}{\underset{\downarrow}{\varepsilon_{obs}}} \qquad \text{(step 1).}$$

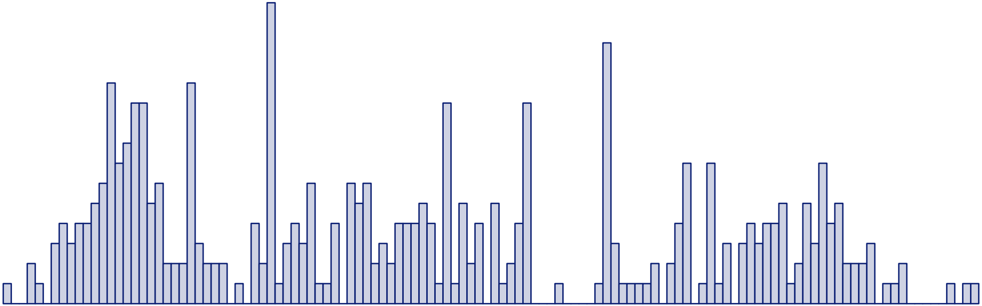
- Imputed values are sampled from $p(\mathbf{y}_{mis} \mid \mathbf{X}_{mis}, \hat{\beta}, \hat{\sigma})$ (step 2).

➡ Step 2 is analogous to step 3 in Bayesian multiple imputation.

Semi-parametric Imputation

Both Bayesian and bootstrap multiple imputation sample imputed values from a distribution $p(\mathbf{y}_{mis} \mid \mathbf{X}_{mis}, \hat{\boldsymbol{\beta}}, \hat{\sigma})$.

Sometimes, the empirical distribution can not be adequately approximated by a known probability distribution.



Semi-parametric Imputation

Predictive Mean Matching (PMM)

- semi-parametric approach to imputation
- developed for settings where the normal distribution is not a good choice for the predictive distribution. (Little, 1988; Rubin, 1986)

Semi-parametric Imputation

Predictive Mean Matching (PMM)

- semi-parametric approach to imputation
- developed for settings where the normal distribution is not a good choice for the predictive distribution. (Little, 1988; Rubin, 1986)

Idea:

- find cases in the observed data that are similar to the cases with missing values
- fill in the missing value with the observed value from one of those cases

To find similar cases, the predicted values of observed and unobserved cases are compared.

Semi-parametric Imputation

The steps in PMM:

1. Obtain parameter estimates for $\hat{\beta}$ and $\hat{\sigma}$ (see later)
2. Calculate the predicted values for the observed cases

$$\hat{\boldsymbol{y}}_{obs} = \mathbf{X}_{obs}\hat{\beta}$$

3. Calculate the predicted value for the missing cases

$$\hat{\boldsymbol{y}}_{mis} = \mathbf{X}_{mis}\hat{\beta}$$

4. For each missing value, find *d* donor candidates that fulfil a given criterion (details on the next slide).
5. Randomly select one of the donors.

Semi-parametric Imputation

Several **criteria to select donors** (donor candidates) have been proposed:

- **Case with the smallest absolute difference** $\left|\hat{y}_{mis,i} - \hat{y}_{obs,j}\right|$, $j = 1, \ldots, q$.

Semi-parametric Imputation

Several **criteria to select donors** (donor candidates) have been proposed:

- **Case with the smallest absolute difference** $\left|\hat{y}_{mis,i} - \hat{y}_{obs,j}\right|, \, j = 1, \ldots, q$.
- Donor candidates are the d **cases with the smallest absolute difference** $\left|\hat{y}_{mis,i} - \hat{y}_{obs,j}\right|, \, j = 1, \ldots, q$. The donor is selected randomly from the candidates.

Semi-parametric Imputation

Several **criteria to select donors** (donor candidates) have been proposed:

- **Case with the smallest absolute difference** $\left|\hat{y}_{mis,i} - \hat{y}_{obs,j}\right|$, $j = 1, \ldots, q$.
- Donor candidates are the d **cases with the smallest absolute difference** $\left|\hat{y}_{mis,i} - \hat{y}_{obs,j}\right|$, $j = 1, \ldots, q$. The donor is selected randomly from the candidates.
- Donor candidates are those cases for which the **absolute difference is smaller than some limit** η: $\left|\hat{y}_{mis,i} - \hat{y}_{obs,j}\right| < \eta$, $j = 1, \ldots, q$. The donor is selected randomly from the candidates.

Semi-parametric Imputation

Several **criteria to select donors** (donor candidates) have been proposed:

- **Case with the smallest absolute difference** $\left|\hat{y}_{mis,i} - \hat{y}_{obs,j}\right|$, $j = 1, \ldots, q$.
- Donor candidates are the d **cases with the smallest absolute difference** $\left|\hat{y}_{mis,i} - \hat{y}_{obs,j}\right|$, $j = 1, \ldots, q$. The donor is selected randomly from the candidates.
- Donor candidates are those cases for which the **absolute difference is smaller than some limit** η: $\left|\hat{y}_{mis,i} - \hat{y}_{obs,j}\right| < \eta$, $j = 1, \ldots, q$. The donor is selected randomly from the candidates.
- Select candidates like in 2. or 3., but select the donor from the candidates with probability that depends on $\left|\hat{y}_{mis,i} - \hat{y}_{obs,j}\right|$.

Semi-parametric Imputation

Potential issues with donor selection

- Selection criteria 2. - 4., **require the number of candidates** *d* (or max. diff. η) to be specified. Common choices for *d* are 3, 5 or 10.

Semi-parametric Imputation

Potential issues with donor selection

- Selection criteria 2. - 4., **require the number of candidates** d (or max. diff. η) to be specified. Common choices for d are 3, 5 or 10.
- If the same donor is chosen in many/all imputations (e.g., because only a few similar observed cases are available), the **uncertainty about the missing values will be underestimated**.

Semi-parametric Imputation

Potential issues with donor selection

- Selection criteria 2. - 4., **require the number of candidates** d (or max. diff. η) to be specified. Common choices for d are 3, 5 or 10.
- If the same donor is chosen in many/all imputations (e.g., because only a few similar observed cases are available), the **uncertainty about the missing values will be underestimated**.
- ➡ PMM may be **problematic** when

Semi-parametric Imputation

Potential issues with donor selection

- Selection criteria 2. - 4., **require the number of candidates** d (or max. diff. η) to be specified. Common choices for d are 3, 5 or 10.
- If the same donor is chosen in many/all imputations (e.g., because only a few similar observed cases are available), the **uncertainty about the missing values will be underestimated**.
- ➡ PMM may be **problematic** when
  - the **dataset is very small**,

Semi-parametric Imputation

Potential issues with donor selection

- Selection criteria 2. - 4., **require the number of candidates** d (or max. diff. η) to be specified. Common choices for d are 3, 5 or 10.
- If the same donor is chosen in many/all imputations (e.g., because only a few similar observed cases are available), the **uncertainty about the missing values will be underestimated**.
- ➡ PMM may be **problematic** when
  - the **dataset is very small**,
  - the **proportion of missing values is large**, or

Semi-parametric Imputation

Potential issues with donor selection

- Selection criteria 2. - 4., **require the number of candidates** d (or max. diff. η) to be specified. Common choices for d are 3, 5 or 10.
- If the same donor is chosen in many/all imputations (e.g., because only a few similar observed cases are available), the **uncertainty about the missing values will be underestimated**.
- ➡ PMM may be **problematic** when
  - the **dataset is very small**,
  - the **proportion of missing values is large**, or
  - **predictor variables** are strongly **related to the missingness**.

Semi-parametric Imputation

Potential issues with donor selection

- Selection criteria 2. - 4., **require the number of candidates** d (or max. diff. η) to be specified. Common choices for d are 3, 5 or 10.
- If the same donor is chosen in many/all imputations (e.g., because only a few similar observed cases are available), the **uncertainty about the missing values will be underestimated**.
- ➡ PMM may be **problematic** when
  - the **dataset is very small**,
  - the **proportion of missing values is large**, or
  - **predictor variables** are strongly **related to the missingness**.
- Using $d = 1$ (selection criterion 1.) is not a good idea. On the other hand, using too many candidates can lead to bad matches.

Semi-parametric Imputation

Potential issues with donor selection

- Selection criteria 2. - 4., **require the number of candidates** *d* (or max. diff. η) to be specified. Common choices for *d* are 3, 5 or 10.
- If the same donor is chosen in many/all imputations (e.g., because only a few similar observed cases are available), the **uncertainty about the missing values will be underestimated**.
- ➡ PMM may be **problematic** when
  - the **dataset is very small**,
  - the **proportion of missing values is large**, or
  - **predictor variables** are strongly **related to the missingness**.
- Using $d = 1$ (selection criterion 1.) is not a good idea. On the other hand, using too many candidates can lead to bad matches.
- Schenker & Taylor (1996) proposed an adaptive procedure to select *d*, but it is not used much in practice.

Semi-parametric Imputation

For the **sampling of the parameters** (step 1), different approaches have been introduced in the literature:

Type-0	$\hat{\beta}_{LS/ML}$ (least squares or maximum likelihood) are used in both prediction models
Type-I	$\hat{\beta}_{LS/ML}$ to predict $\hat{y}_{obs}$; $\tilde{\beta}_{B/BS}$ (Bayesian or bootstrapped) to predict $\hat{y}_{mis}$
Type-II	$\tilde{\beta}_{B/BS}$ to predict $\hat{y}_{obs}$ as well as $\hat{y}_{mis}$
Type-III	different draws $\tilde{\beta}^{(1)}_{B/BS}$ and $\tilde{\beta}^{(2)}_{B/BS}$ to predict $\hat{y}_{obs}$ and $\hat{y}_{mis}$, respectively

The use of Type-0 and Type-I matching **underestimates the uncertainty** about the regression parameters.

Semi-parametric Imputation

Another point to consider:
the **choice of the set of data used to train the prediction models**

By default, the same set of data (all cases with observed y) is used to train the model and to produce predicted values of y_{obs}.

The predictive model will likely fit the observed cases better than the missing cases, and, hence, **variation will be underestimated**.

Alternatives:

- the **model could be trained on the whole data** (using previously imputed values)
- use a **leave-one-out approach** on the observed data

References I

Little, R. J. (1988). Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics*, *6*(3), 287–296.

Morris, T. P., White, I. R., & Royston, P. (2014). Tuning multiple imputation by predictive mean matching and local residual draws. *BMC Medical Research Methodology*, *14*(1), 75.

Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics*, *4*(1), 87–94.

Schenker, N., & Taylor, J. M. (1996). Partially parametric techniques for multiple imputation. *Computational Statistics & Data Analysis*, *22*(4), 425–446.

References II

Siddique, J., & Belin, T. R. (2008). Multiple imputation using an iterative hot-deck with distance-based donor selection. *Statistics in Medicine*, *27*(1), 83–102.