# Multiple Imputation for Longitudinal/Hierarchical Data

**Mike Kenward**

**James Carpenter**

**London School of Hygiene and Tropical Medicine**

Funded in part by the UK ESRC Research Methods Programme

## Contents:

1. Introduction, multiple imputation
2. Imputation methods for simple cross-sectional data
3. Imputation methods for ‘structured’ (multilevel, hierarchical, longitudinal) data
4. An implementation in MLwiN
5. Sensitivity analysis

# Some notation and definitions

*All response/outcome data*, whether observed or not:

$$\mathbf{Y} = \{\mathbf{Y}_O, \mathbf{Y}_M\}.$$

$\mathbf{Y}_O$: observed, $\mathbf{Y}_M$: missing.

*Covariates:*

$$\mathbf{X} = \{\mathbf{X}_O, \mathbf{X}_M\}.$$

$\mathbf{X}_O$: observed, $\mathbf{X}_M$: missing.

Depending on the context these may all refer to one unit, or to an entire dataset.

Define

**Z**= {**Y**,**X**}, $\mathbf{Z}_O = \{\mathbf{Y}_O, \mathbf{X}_O\}$, and $\mathbf{Z}_M = \{\mathbf{Y}_M, \mathbf{X}_M\}$

*Missing value indicator:*

Corresponding to every element of **Z**, there is an $R$:

$$R = \begin{cases} 1 & \text{if observed} \\ 0 & \text{if missing} \end{cases}$$

with **R** $= \{R\}$.

**Goal:** make inferences about

$$f(\mathbf{Y} \mid \mathbf{X})$$

using $\mathbf{Y}_O$ and $\mathbf{X}_O$ (and $\mathbf{R}$).

How are these connected?

$$f(\mathbf{Y}_O \mid \mathbf{X}_O) = \int\int f(\mathbf{Y}_O, \mathbf{Y}_M \mid \mathbf{X}_O, \mathbf{X}_M) f(\mathbf{X}_M \mid \mathbf{X}_O) \mathbf{dX}_M \mathbf{dY}_M.$$

## **Estimation**:

Suppose we have an unbiased (vector valued) estimating equation for the complete data

$$\mathbf{U}(\mathbf{Y};\mathbf{X};\hat{\boldsymbol{\theta}}) = \sum_{i=1}^{n} \mathbf{U}_i(\mathbf{Y}_i;\mathbf{X}_i;\hat{\boldsymbol{\theta}}) = \mathbf{0}.$$

then with missing responses the following estimating equation remains unbiased (hence consistent for $\boldsymbol{\theta}$)

$$\sum_{i=1}^{n} \{R_i\mathbf{U}_i + (1 - R_i)\mathsf{E}_g(\mathbf{U}_i)\} = \mathbf{0}.$$

where the expectation is taken over the joint distribution of the missing data $\mathbf{Z}_M$ conditional on the observed data (the conditional predictive distribution $g(\mathbf{Z}_M)$).

This will be the score equation in the likelihood setting.

Under MAR we do *not* need to condition on **R**.

There is an alternative method that does not require knowledge of $g(\mathbf{Z}_M)$.

If we know (or can estimate) the probability of observing a complete unit, say $\pi_i$, then the following weighted estimating equations are unbiased:

$$\sum_{i=1}^{n} \frac{R_i \mathbf{U}_i}{\pi_i} = \mathbf{0}.$$

So broadly there are two routes:

1. Correct the estimating equations using

   $$\mathrm{E}_g(\mathbf{U}_i)$$

   This requires distributional information, but is the more precise.

   [Direct empirical (likelihood) versions are called **mean score** methods.]

2. Use the inverse of $\pi_i$ as a weight (*e.g.* Horvitz-Thompson).

   More robust but less precise in its simple form, but more efficient versions have been proposed.

- There is a sense nearly all principled approaches to the problem missing values are variations on these two themes.

  [There is a connection between them through the nonparametric (hot-deck) estimation of of $g(\mathbf{Z})$]

- In practice, key distributions, or probabilities of missingness, will be unknown.

- Untestable assumptions will be made in estimating these.

- Its important to distinguish between assumptions/models and technology for fitting and managing these (*e.g.* likelihood, full Bayes, pseudo-likelihood, hybrid MCMC methods, weighted estimating equations, multiple imputation,. . . )

# Multiple Imputation

*Some References*

- Gilks, W.R., Richardson, S. & Spiegelhalter, D.J. (1996) *Markov Chain Monte Carlo in Practice.* Chapman & Hall/CRC.
- Rubin, D.B. (1987) *Multiple Imputation for Nonresponse in Surveys.* Wiley.
- Rubin, D.B. (1996) Multiple imutation after 18+ years. *J. Amer. Statist. Ass.*, **91**, 473–489.
- Schafer, J.L. (1997) *Analysis of Incomplete Multivariate Data.* Chapman & Hall/CRC.

## An intuitive view

Starting point:

- Analysis of a complete data set is relatively simple.
- By comparison, estimation and inference with missing data is awkward.
- The conditional predictive distribution can be estimated from the observed data (true under MAR).

Consider the situation with parameter $\theta$.

We take a Bayesian approach.

The posterior is given by

$$\begin{aligned} f(\boldsymbol{\theta} \mid \mathbf{Z}_O) &= \int f(\boldsymbol{\theta} \mid \mathbf{Z}_O, \mathbf{Z}_M) f(\mathbf{Z}_M \mid \mathbf{Z}_O) d\mathbf{Z}_M \\ &= E_g\{f(\boldsymbol{\theta} \mid \mathbf{Z}_O, \mathbf{Z}_M)\}, \end{aligned}$$

with posterior mean

$$\begin{aligned} \mathrm{E}(\boldsymbol{\theta} \mid \mathbf{Z}_O) &= \mathrm{E}_g[\mathrm{E}\{\tilde{\boldsymbol{\theta}}(\mathbf{Z}_O, \mathbf{Z}_M)\}] \\ &\approx \frac{1}{M} \sum_{i=1}^{M} \tilde{\boldsymbol{\theta}}(\mathbf{Z}_O, \mathbf{Z}_M^i) \end{aligned}$$

where $\mathbf{Z}_M^i$ is drawn from $g(\cdot)$, and $\tilde{\boldsymbol{\theta}}(\mathbf{Z}_O, \mathbf{Z}_M^i)$ is the mean of the posterior given $\mathbf{Z}_O$ and $\mathbf{Z}_M^i$.

Similarly the posterior covariance matrix reduces to

$$\begin{aligned}\mathsf{V} &= \mathsf{E}_g\{\mathsf{V}(\boldsymbol{\theta} \mid \mathbf{Z}_O, \mathbf{Z}_M)\} + \mathsf{V}_g\,\{\mathsf{E}(\boldsymbol{\theta} \mid \mathbf{Z}_O, \mathbf{Z}_M)\} \\ &\approx \frac{1}{M}\sum_{i=1}^{M} \mathsf{V}\{\tilde{\boldsymbol{\theta}}(\mathbf{Z}_O, \mathbf{Z}_M^i)\} \\ &+ \left(1 + \frac{1}{M}\right)\left(\frac{1}{M-1}\right)[\sum_{i=1}^{M}\{\tilde{\boldsymbol{\theta}}_{MI} - \tilde{\boldsymbol{\theta}}(\mathbf{Z}_O, \mathbf{Z}_M^i)\}\{\tilde{\boldsymbol{\theta}}_{MI} - \tilde{\boldsymbol{\theta}}(\mathbf{Z}_O, \mathbf{Z}_M^i)\}^T].\end{aligned}$$

for

$$\tilde{\boldsymbol{\theta}}_{MI} = \frac{1}{M}\sum_{i=1}^{M} \tilde{\boldsymbol{\theta}}(\mathbf{Z}_O, \mathbf{Z}_M^i).$$

In practice the response model is fitted to each of $M$ completed datasets.

The posterior means $\tilde{\boldsymbol{\theta}}(\mathbf{Z}_O, \mathbf{Z}_M^i)$ are replaced the maximum likelihood estimates.

And the posterior covariance matrices $\mathsf{V}(\tilde{\boldsymbol{\theta}} \mid \mathbf{Z}_O, \mathbf{Z}_M)$ are estimated using the inverse information.

So, for a single parameter MI looks like this:

For each draw from $g(\cdot)$ solve the score to get the ML estimator

$\tilde{\theta}^i$ and variance (inverted information) $\tilde{V}^i$.

These are then combined:

$$\begin{aligned}
\tilde{\theta} &= \frac{1}{M}\sum_{i=1}^{M}\tilde{\theta}^i \\
\tilde{V}(\tilde{\theta}) &= \frac{1}{M}\sum_{i=1}^{M}\tilde{V}^i + \frac{1+M^{-1}}{M-1}\sum_{i=1}^{M}\left[\tilde{\theta}^i - \tilde{\theta}\right]^2 .
\end{aligned}$$

Typically $M$ can be small (say 5-10).

What are the conditions for appropriate (proper) imputations?

Rubin gives formal rules (1987, ch 4).

In practice these are almost never checked formally, and the following guidelines form the basis for the justification of the many various procedures used.

Rubin (1987, pp.126–127):

- “Draw imputations following the Bayesian paradigm as repetitions from a Bayesian posterior distribution of the missing values under the chosen models for nonresponse and data, or an approximation to this posterior distribution that incorporates appropriate between-imputation variability.” [*proper* imputations]
- ”Choose models of nonresponse appropriate for the posited response mechanism.”
- “Choose models for the data that are appropriate for the complete-data statistics likely to be used - if the model for the data is correct, then the model is appropriate for all complete-data statistics.”

## Imputer and Analyst

- MI originated in the sample survey setting.
- Many different analyses may be performed from one multiply imputed dataset.
- The “imputer’s” model need not be the same as the “analyst’s” model, **but** if the latter contains structure that the former does not there may be serious problems.
  [Fay, R.E. (1992) *Proc. Survey Res. Meth. Sec. Amer. Statistic. Ass.*, 227–232.]
- For a strictly Bayesian interpretation, the imputer’s and analysts’s models must coincide.

- Obviously a key part of the procedure is the formulation and use of the conditional predictive distribution.

- When only *responses* are missing, *e.g.* longitudinal clinical trials, direct likelihood based modelling (Bayesian or frequentist) is usually more straightforward and transparent.

  In practice, typically, MI just approximates the inferences produced by such approaches.

When we have missing covariates, and a well defined joint distribution (e.g. multivariate normal) direct *estimation* may still be relatively easy: *e.g.* obtain ML estimates of the parameters of the conditional distribution

$$f(\mathbf{Y} \mid \mathbf{X})$$

from those of the the joint distribution

$$f(\mathbf{Y}, \mathbf{X}_M \mid \mathbf{X}_O).$$

But what about precision?

The conditional framework is awkward: but is accommodated by the variance formula in MI.

## Do imputations have to incorporate the uncertainty in the model parameters?

- It's not obvious why it is wrong to impute from

  $$\mathbf{Z}_M \mid \mathbf{Z}_O(, \mathbf{R}).$$

  with parameters *fixed* at appropriate values (consistent estimators).

- In a sense there is nothing wrong with it, this approach *can* produce consistent estimators for the substantive model: but estimates of precision are not as simple as with MI.

For a rigorous discussion of proper and improper imputation, and variance estimation, with these types of stochastic imputation procedures, see

Wang N and Robins JM (1998) Large-sample thoery for parametric multiple imputation procedures. *Biometrika*, **85**, 935–948.

Robins JM and Wang N (2000) Inference for imputation estimators. *Biometrika*, **85**, 113–124.

## Conclusions for our setting:

- With missing outcomes only - use direct modelling approaches (likelihood, estimating equations...)
- For smaller problems where covariates are missing, do formal Bayesian analyses using WinBugs.
- For larger problems with missing covariates (the majority) MI provides a practical *general* basis for calculating variances and subsequent inference for regression models.

  An additional advantage is the separation of the substantive and imputation model.

# Imputation methods for simple cross-sectional data

(Unsurprisingly) much of the work in MI surrounds the choice of, and sampling from, the **conditional predictive distribution:**

$$g(\mathbf{Z}_M) = f(\mathbf{Y}_M, \mathbf{X}_M \mid \mathbf{Y}_O, \mathbf{X}_O, \mathbf{R}).$$

MAR is nearly always assumed meaning that we can use

$$g(\mathbf{Z}_M) = f(\mathbf{Y}_M, \mathbf{X}_M \mid \mathbf{Y}_O, \mathbf{X}_O).$$

It has been suggested that inferences are fairly robust to this (*e.g.* Schafer, 2000).

In sufficiently large samples we can often do acceptably well by approximating the posterior predictive distribution using a multivariate normal with mean and covariance matrix taken from the maximum likelihood estimates.

MCMC methods can (and are) also used to obtain draws from the appropriate posterior(s).

In real problems we will usually be faced with the problem of imputing among a set of mixed variable types, for both response and covariate: $\{\mathbf{Y}_M, \mathbf{X}_M \mid \mathbf{Y}_O, \mathbf{X}_O\}$

- continuous
- binary
- categorical: ordinal
- categorical: nominal

Joint modelling of such variables in a general, flexible, way is far from trivial.

There are two main classes of approach in MI, both involve approximation.

- (I) Use a convenient class of multivariate model as an approximation. This usually implies:
  - multivariate normal (or Gaussian) for continuous, ordinal and binary variables,
  - a loglinear model for nominal categorical
- (II) Use univariate conditional models in a Gibbs sampler *type* approach.

## (II) Gibbs type imputation procedures

- While it can be difficult to construct an appropriate joint imputation model without using gross approximations, it is much simpler to model (and hence impute) each variable separately conditional on the others.

- Indeed, this is precisely the problem often met in Bayesian methods when sampling from the full joint posterior.

  Bayesian applications have been transformed by the use of the **Gibbs sampler.**

An imputation approach that applies the spirit of Gibbs sampling in the MI setting is as follows.

- We want to impute from the joint posterior distribution of $\{Z_1, \ldots, Z_P\}$.

- Impute instead, in turn, from the (approximate) conditional posteriors

$$z_j \mid z_1, \ldots, z_{j-1}, z_{j+1}, \ldots, z_P, \quad j = 1, \ldots, P.$$

  where missing values among the conditioning variables are replaced by their previous imputations.

- Repeat the process T times.

- Use the last set of imputations to complete one MI data set.

A similar idea from two different groups, principally

*The Sequential Regression Imputation Method*

Ragnuthan TE, Lepkowski J, van Hoewyk, J and Solenberger PW (2001) A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*.

with associated software: **IVEWARE**

`www.isr.urmich.edu/src/smp/ive`.

An example:

Taylor MG, Cooper KL, Wei JT, Sarma AV, Raghunathan TE, Heeringa SG (2002) use of multiple imputation to correct for nonresponse bias in a survey of urologic symptoms among African-American men. *American Journal of Epidemiology*, **156**, 774–782.

## *Multivariate Imputation by Chained Equations*

Van Buuren S, Boshuizen HC and Knook DL (1999) Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, **18**, 681–694.

with associated S+ software package **MICE**

`www.multiple-imputation.com`

There is now a version of MICE implemented in Stata.

[Royston P (2004) Multiple imputation of missing values. *The Stata Journal*, **4**, 227-241.]

`www.stata.com/support`

What about the justification for these methods?

The existence of a joint limiting distribution is not guaranteed.

To quote van Buuren and Oudshoorn (MICE):

“It is hard to establish convergence in the general case, but simulation studies suggest that the coverage properties in some important practical cases are quite good.”

Gelman and Raghunathan (2001, Statist. Sci. 268–269)

“..the study of conditional distributions is an area where theory has not caught up with practice.”

We know this works for multivariate normal settings.

With monotone missing data patterns and sequential imputation (from most to least incomplete) the joint distribution is well defined.

Simulations studies by both groups (and some of our own) suggest it works well more generally.

A big advantage is the ability to accommodate restrictions and bounds on particular variables.

## Imputation methods for ‘structured’ data

- We are dealing with the problem of data from a hierarchial structure.
- Additional longitudinal structure will be common, not necessarily with times of measurement common to all subjects.

  (Simple attrition in a wave based survey is easier to deal with.)
- Some covariates will have missing values, possibly at different levels of the hierarchy.

Again, we might consider two approaches:

1. (I) Full joint modelling: need a framework for ‘structured’ multivariate data of different types.
2. (II) Gibbs type approach.

Look at (II) first

Now each individual conditional model will be ‘structured’, for example for

$$\mathbf{Z}^{(i)} = \{Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_P\}$$

- Continuous: general linear mixed model

$$\mathbf{Z}_i \mid \mathbf{z}^{(i)}, \mathbf{u} \sim \mathrm{N}(\mathbf{x}_o\boldsymbol{\beta} + \mathbf{z}^{(i)}\boldsymbol{\Gamma} + \mathbf{u}\mathbf{H}; \quad \boldsymbol{\Delta})$$

- Binary: generalized linear mixed model

$$\mathrm{logit}\{\mathrm{P}(Z_{ij} = 1 \mid Z_j^{(i)}, \mathbf{u})\} = \mathbf{x}_o\boldsymbol{\beta} + Z_j^{(i)}\boldsymbol{\Gamma} + \mathbf{u}\mathbf{H}$$

Similar for ordinal.

IVEWARE accommodates complex sample design using survey based methods for estimation of precision.

We have taken the first route however with well defined joint posteriors, using and developing the facilities already available in MLwiN.

- All variables continuous: we can fit multivariate ‘structured’ data using existing tools.
- We use an unstructured covariance matrix across variables, which can be combined with appropriate random effects (and other implied) structures for the hierarchical/longitudinal within-variable component.

Finally, MLwiN has an existing MCMC tool that allows draws from the posterior for a fitted multivariate model (with appropriate priors).

References:

Rasbash, Steele, Browne and Prosser (2004) *A user's guide to* `MLwiN` *(version 2.0)*, London: Institute of Education.

Browne (2003) *MCMC estimation in* `MLwiN` *version 2.0* , London: Institute of Education.

Schafer's standalone package PAN (for Windows) has similar facilities for multivariate normal data (`www.multiple-imputation.com`)

We are developing an MLwiN macro that will, for a general multilevel model, do the following.

- Takes a chosen imputation model $g(\mathbf{Z}_M)$ (under the MAR assumption) and fits this using ML.

  There may be additional covariates that do not appear in the response model.

- Uses the MCMC tool to draw the required number of imputations (allowing adequate burn-in and gaps between draws).

- Fits the response model to each imputed set, combines the results using Rubin's rules and calculate tests/CI's.

The macro:

- takes the defined response model $\mathbf{Y} \mid \mathbf{X}$;
- identifies all incomplete covariates;
- sets up a multivariate imputation model, $\mathbf{Z}_M \mid \mathbf{Z}_O$
- sets up a default covariance structure for the imputation model based on the structure of the response model;
- uses MCMC to draw from the posterior of the imputation model.

Points to note:

- The MCMC tool needs one complete response: the macro chooses this and adds it to the multivariate model.

- The user can modify the imputation model.

- Only multivariate normal at present.

# A simple example: the Class-Size Study

[Blatchford *et al.* *Brit. J. Educ. Res.* (2002)]

| | |
|---|---|
| uniqueid: | Unique pupil identifier |
| schn: | School identifier |
| year: | Year: reception (year=1) or first year (year=2). |
| litbase: | Baseline literacy score (either pre-reception or pre-year 1, if pupil has no reception year data) |
| nlit: | Literacy score at the end of the reception or first year (depending on value of year) |
| nmat: | Maths score at the end of the reception or first year (depending on the value of year) |
| csize: | Size of class |
| fsmn: | Eligible for free school meals (1=yes, 0=no) |
| gend: | Sex (1=boys, 0=girls) |
| tentry: | Term of school entry (1=Spring or Summer, 0=Autumn) |
| cons: | Constant, set to 1. |

## Missing Values:

```
nlit:   105
nmat:   122
csize:  471
```

Analysis restricted to only those with complete data uses only 848 out of 1408 observations

*Response Model*:

$i$: individual, $j$: school

$$\begin{aligned}\text{nlit}_{ij} = \beta_0 &+& \beta_1\text{litbase}_{ij} + \beta_2\text{csize}_{ij} + \beta_3\text{year}_{jk} + \beta_4\text{fsmn}_{ij} \\ &+& \beta_5\text{gender}_{ij} + \beta_6\text{tentry}_{ij} + e_j + e_{ij}\end{aligned}$$

$$\begin{aligned}e_j &\sim& \text{N}(0, \sigma_s^2) \\ e_{ij} &\sim& \text{N}(0, \sigma^2)\end{aligned}$$

*Imputation Model*:

$$\begin{aligned}
\text{nlit}_{jk} = \beta^*_{12} &+ \beta^*_0\text{year}_{jk} + \beta^*_3\text{fsmn}_{jk} + \beta^*_6\text{gender}_{jk} + \beta^*_9\text{tentry}_{jk} \\
&+ e_{1j} + e_{1jk} \\
\text{litbase}_{jk} = \beta^*_{13} &+ \beta^*_1\text{year}_{jk} + \beta^*_4\text{fsmn}_{jk} + \beta^*_7\text{gender}_{jk} + \beta^*_{10}\text{tentry}_{jk} \\
&+ e_{2j} + e_{2jk} \\
\text{csize}_{jk} = \beta^*_{14} &+ \beta^*_2\text{year}_{jk} + \beta^*_5\text{fsmn}_{jk} + \beta^*_8\text{gender}_{jk} + \beta^*_{11}\text{tentry}_{jk} \\
&+ e_{3j} + e_{3jk}
\end{aligned}$$

$$\begin{pmatrix} e_{1j} \\ e_{2j} \\ e_{3j} \end{pmatrix} \sim \text{N}(\mathbf{0}, \mathbf{\Sigma}_s), \quad \begin{pmatrix} e_{1jk} \\ e_{2jk} \\ e_{3jk} \end{pmatrix} \sim \text{N}(\mathbf{0}, \mathbf{\Sigma}_y)$$

# Extensions: other types of variable.

Use an unstructured multivariate set of underlying (latent) variables.

- Binary/ordinal.

  Use a probit link.

- Nominal variables:

  Broadly follow the approach of

  Albert JH and Chib S (1993) Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, **88**, 669–679.

## Sensitivity analysis

The relative simplicity of the MI imputations follows from the MAR assumption.

The MI route does allow a relatively simple way of assessing sensitivity to this assumption:

modify the imputation model to allow for explicit non-random dependencies.

(Because we generate the imputations we can allow nonresponse to depend on ’missing values’.)

In the current setting we can make this modified nonresponse procedure very explicit by using a postulated accept-reject mechanism when the imputations are drawn.

This has suggested (and done) by several authors in particular settings.

This can be inefficient and we are proposing using a weighted method analogous to that used for obtaining posteriors in some settings.

It is planned to add this facility to the MLwiN macro.

*Macro References:*

Carpenter and Goldstein (2004) Multilevel Imputation in MLwiN:

`http://www.lshtm.ac.uk/msu/missingdata/papers/newsletterdec04.pdf`

The macro can be downloaded from

`http://www.lshtm.ac.uk/msu/missingdata/software.html`