Introduction to Multiple Imputation

James Carpenter & Mike Kenward

Department of Medical Statistics
London School of Hygiene & Tropical Medicine

James.Carpenter@lshtm.ac.uk
https://missingdata.lshtm.ac.uk

June 2005

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Table of Contents

Introduction

The aim of this presentation is to:

1. introduce the ideas of multiple imputation;
2. outline how to carry out multiple imputation, and
3. provide an intuitive justification for multiple imputation.

Why do multiple imputation?

One of the main problems with the single stochastic imputation methods is the need for developing appropriate variance formulae for each different setting.

Multiple imputation attempts to provide a procedure that can get the appropriate measures of precision relatively simply in (almost) any setting.

It was developed by Rubin is a survey setting (where it feels very natural) but has more recently been used more widely.

Below, we assume we have an established method for fitting our model, had the data been completely observed.
- e.g. regression, glm, ...

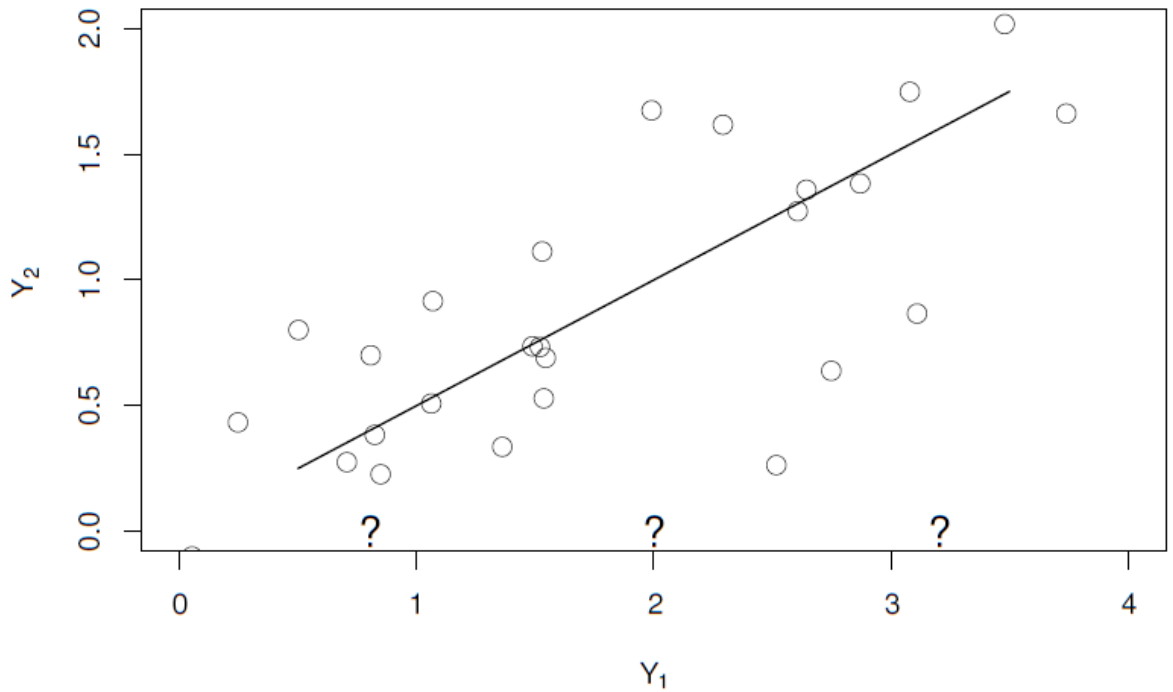
For simplicity, suppose we have only two variables in our data set.
Suppose one of them is observed on every unit. Call this Y_1.
Suppose one is only observed on some units. Call this Y_2.

The key idea

The key idea is to use the data from units where both (Y_1, Y_2) are observed to learn about the relationship between Y_1 and Y_2. Then, we use this relationship to complete the data set by drawing the missing observations from the distribution of $Y_2|Y_1$. We do this K (typically 5) times, giving rise to K complete data sets.
We analyse each of these data sets in the usual way.
We combine the results using particular rules.

Intuition behind multiple imputation

First, we model observed (Y_1, Y_2) pairs. These are shown below, with a regression line through them. It's crucial that the variable with the missing values is the response, whether or not it is going to be the response in the final model of interest. The '?' indicates we have the value of Y_1, but that for Y_2 is missing.



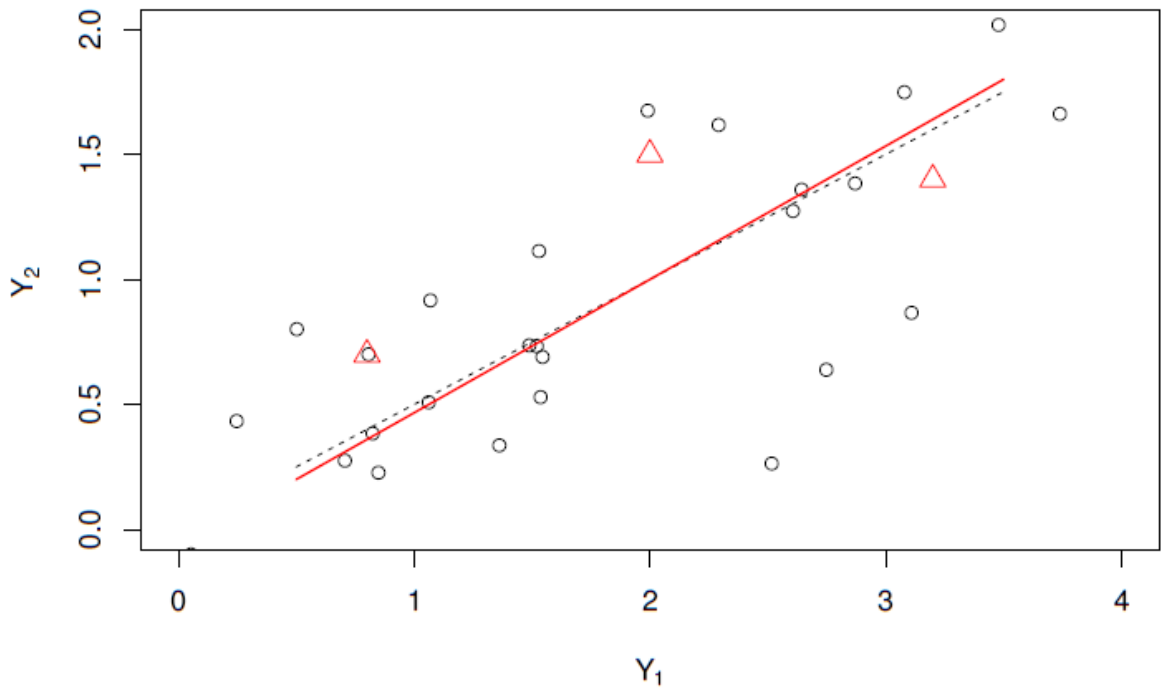
LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Intuition behind multiple imputation

Next, we draw missing Y_2 by (i) drawing from distribution of regression line (ii) drawing from variablity about that line. In the picture below, the dotted line is the regression line from the observed data (as on the previous picture) and the red line is drawn from the estimated distribution of the regression line (i.e. the red line's intercept and slope are drawn from the estimated bivariate normal distribution of the intercept and slope).

Intuition behind multiple imputation

Then, a draw is made from the estimated normal distribution of the residuals, and added to the line, to give the imputed points, shown by red triangles.



LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Intuition behind multiple imputation

From this graph we can see straight away why replacing the missing observations with the mean of Y_2 is a bad idea. For instance, the leftmost '?' in the fist picture above would be given a value far above the regression line (which represents its expected value given Y_1).

We can also see why a single imputation on the regression line - i.e. where the imputed data (triangles in the graph above) lies on the regression line - is inadequate. This would be an over-confident prediction of the missing value. Systematically doing this would lead to estimates of standard errors that were too small, and inferences that were therefore over-confident.

Intuition behind multiple imputation

However, a single imputation of each missing value is not adequate, because we only know the distribution of the missing values. Thus, we need to repeat the imputation process a number of times, each time drawing a new regression line, and new residuals about that regression line. We thus end up with a number of completed data sets as follows:

	Data		Imputation 1		Imputation 2		Imputation 3		Imputation 4	
Unit	Y_M	Y_O	Y_M	Y_O	Y_M	Y_O	Y_M	Y_O	Y_M	Y_O
1	1.1	3.4	1.1	3.4	1.1	3.4	1.1	3.4	1.1	3.4
2	1.5	3.9	1.5	3.9	1.5	3.9	1.5	3.9	1.5	3.9
3	2.3	2.6	2.3	2.6	2.3	2.6	2.3	2.6	2.3	2.6
4	3.6	1.9	3.6	1.9	3.6	1.9	3.6	1.9	3.6	1.9
5	0.8	2.2	0.8	2.2	0.8	2.2	0.8	2.2	0.8	2.2
6	3.6	3.3	3.6	3.3	3.6	3.3	3.6	3.3	3.6	3.3
7	3.8	1.7	3.8	1.7	3.8	1.7	3.8	1.7	3.8	1.7
8	?	0.8	**0.2**	0.8	**0.8**	0.8	**0.3**	0.8	**2.3**	0.8
9	?	2.0	**1.7**	2.0	**2.4**	2.0	**1.8**	2.0	**3.5**	2.0
10	?	3.2	**2.7**	3.2	**2.5**	3.2	**1.0**	3.2	**1.7**	3.2

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Notation for analyses of imputed data sets

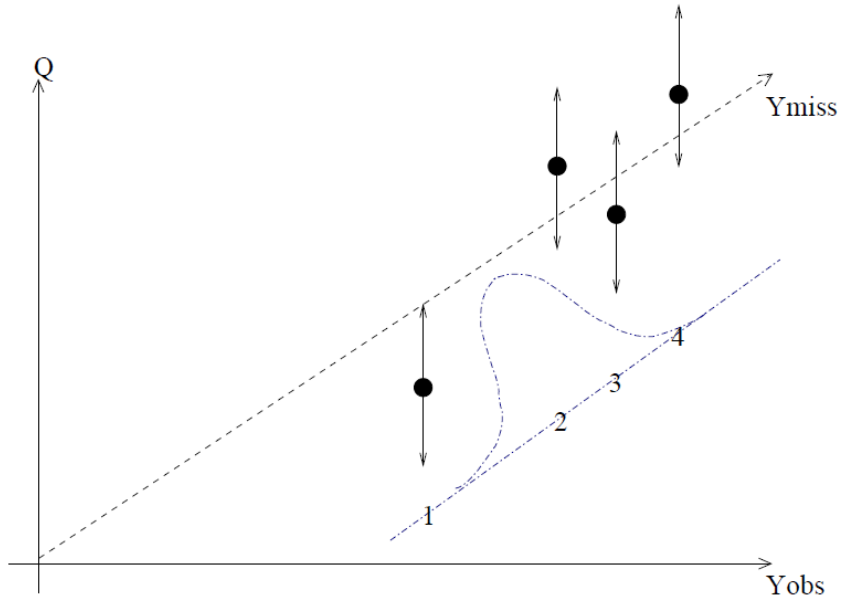
As described above, we have imputed K complete data sets. Analysing each of them in the usual way (i.e. using the model intended for the complete data) gives us K estimates of the original quantity of interest, Q. Denote these estimates $Q_1, \ldots, Q_K$. So, each Q could represent a regression coefficient from a regression model of interest which we fit to each imputed data set in turn.

The analysis of each imputed data set will also give an estimate of the variance of Q_k, say σ_k^2. Again, this is the usual variance estimate from the model.

We combine these quantities to get our overall estimate and its variance using certain rules.

Intuition for combining the estimates

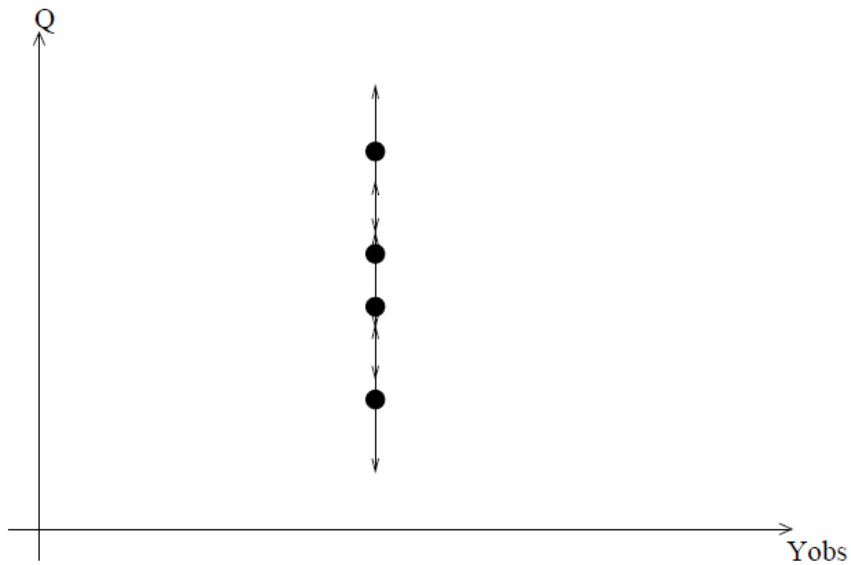
Consider the imputation of just 1 missing observation.



Imagine a 3-d representation, with the Y_{miss} axis going back into the screen. Given a particular value of Y_{obs} the imputations (numbered 1,2,3,4) combine with the observed data to give the estimates of Q shown by the black dots. Each of these estimates also has a variance, which is represented by the line through the black dot.

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Intuition for combining the estimates



Now we project this into two-dimensions, over Y_{miss}.

Intuition for combining the estimates

The multiple imputation estimate is going to be the average of the black dots. In other words, the average over the distribution of Y_M given Y_O of Q, which is itself calculated from the observed and "missing" data:

$$Q_{MI} = \mathbf{E}_{Y_M|Y_O}\mathbf{E}[Q(Y_O, Y_M)]. \quad (1)$$

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Intuition for combining the estimates

The variance has to reflect two components; the variance of the Q's from the imputed datasets about their average and also the variance of each Q estimate. In fact, it is the sum of these two; i.e. in this case (with $Q_1, \ldots, Q_4$) the sample variability of $Q_1, ..., Q_4$ about their mean, plus the average of the variance of $Q_1, ..., Q_4$. These are known respectively as the between imputation variance and the within imputation variance.

Mathematically,

$$\mathbf{V}[Q_{MI}] = \mathbf{E}_{Y_M|Y_O}\mathbf{V}[Q(Y_O, Y_M)] + \mathbf{V}_{Y_M|Y_O}\mathbf{E}[Q(Y_O, Y_M)].$$

This motivates the formulae for combining the estimates and calculating the variance, which are given in the next section.

Combining the estimates

Let the multiple imputation estimate of Q be Q_{MI}. Then, following from the above,

$$Q_{MI} = \frac{1}{K} \sum_{k=1}^{K} Q_k.$$

Further define the within imputation and between imputation components of variance by

$$\sigma_w^2 = \frac{1}{K} \sum_{i=1}^{K} \sigma_k^2, \quad \text{and} \quad \sigma_b^2 = \frac{1}{K-1} \sum_{k=1}^{K} (Q_k - Q_{MI})^2,$$

where we recall our definition of $\hat{\sigma}_k^2 = \mathbf{V}[Q_k]$. Then the variance of Q_{MI} is

$$\sigma_{MI}^2 = \left(1 + \frac{1}{K}\right) \sigma_b^2 + \sigma_w^2.$$

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Testing hypotheses

We assume that, if the data were all observed, then our estimator Q would have a normal distribution.
If this is so, we can compare

$$\frac{Q_{MI} - Q}{\sigma_{MI}} \sim t_\nu,$$

a t-distribution with ν degrees of freedom, where

$$\nu = (K-1)\left[1 + \frac{\sigma_w^2}{(1+1/K)\sigma_b^2}\right]^2.$$

Testing hypotheses

The rate of missing information

If there were no missing data, and we used multiple imputation, we should find that $(1 + 1/K)\sigma_b^2 = 0$. Thus the relative increase in variance due to the missing data is

$$r = \frac{(1 + 1/K)\sigma_b^2}{\sigma_w^2}.$$

Alternatively, the 'rate of missing information' is

$$\frac{(1 + 1/K)\sigma_b^2}{\sigma_w^2 + (1 + 1/K)\sigma_b^2} = \frac{r}{1 + r}.$$

It turns out a better estimate of this quantity is

$$\lambda = \frac{r + 2/(\nu + 3)}{1 + r}.$$

How do we draw $Y_M|Y_O$?

In the pictures above, we described a regression method for drawing $Y_M|Y_O$. This should work reasonably if the data set is large, as it is then an approximation to a Bayesian rule.
This rule says that, if θ is the parameter describing the joint distribution of (Y_O, Y_M) :

Posterior dist^n of (Y_M, θ) given Y_O	$\propto$	Joint dist^n of (Y_M, Y_O) given θ	X	dist^n of θ.

We put an uninformative distribution on θ, and discard the values of θ drawn from the posterior, leaving a sample from $Y_M|Y_O$.

Frequently asked questions

- How many imputations?
  - With 50% missing information, an estimate based on 5 impuitations has SD 5% wider than one with an infinite number of impuations.
- What if not MAR?
  - Most software implementations assume MAR, but this is not necessary.
- Why not compute just one imputation?
  - Underestimates variance, as can't estimate $\hat{\sigma}_b^2$.
- What if I am interested in more than one parameter?
  - Imputation proceeds in the same way, as does finding the overall estimate of Q. However, the estimating the covariance matrix can be tricky. Typically more imputations will be needed. See Schafer (2000) for a discussion.

Some references for MI

Shafer (1999): Overview of how you do MI

Schafer (1997): Key book giving details of data augmentation and MI methods in many models.

Rubin (1976): Article bringing together the theory in an accessible way (for mathematical statisticians).

Rubin (1996): review of the use of MI after $\sim$ 18 years.

Horton and Lipsitz (2001): Comparison of software packages.

Allison (2000): - a cautionary tale!

Software for drawing $Y_M|Y_O$.

We can use Markov Chain Monte Carlo (MCMC) methods to draw from this posterior distribution, and then we discard the θ's and use the Y_M's as described above.

This approach is implemented in *MLwiN*. Other options include WinBUGS or PROC MI in SAS.
Note that drawing from $Y_M|Y_O$ and then doing the analysis in WinBUGS can be unfeasibly slow even for moderate data sets.

One alternative is to use 'chained equations' also known as 'regression switching' or 'sequential regression imputation' (all variants of the same approach).
Please see the software page on our website.

Chained equations: some comments

Roughly, multiple imputation using chained equations proceeds as follows. (We say 'roughly', as implementations vary):

1. To get started, for each variable in turn fill in missing values with randomly chosen observed values
2. 'filled-in' values in the first variable are discarded leaving the original missing values. These missing values are then imputed using regression imputation on all other variables.
3. The 'filled-in' values in the second variable are discarded. These missing values are then imputed using 'proper' regression imputation on all other variables.
4. This process is repeated for each variable in turn. Once each variable has been imputed using the regression method we have completed one 'cycle'.
5. The process is continued for several cycles, typically ~ 10.

Chained equations: some comments

This was first published by Raghunathan et. a. (2001); see also the SAS implementation.

For a medical example see Taylor et. al. (2002).

A Dutch group has developed related software; see Van Buuren et. al. (1999), and associated S+ software.

This has been implemented in `stata`; see Royston (2004), and stata help pages.
All the implementations are slightly different!

Although MICE is an attractive approach, overcoming some of the issues with binary and ordinal data that are difficult for proper multiple imputation, the lack of a well established theoretical basis means even those who propose it suggest it is used cautiously.

Chained equations: some comments

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.'

Chained equations: some comments

The problem is that you are in effect defining many conditional distributions, and this does not guarantee the existence of a joint distribution.

Further discussion is given by Raghunathan et. a. (2001) (the original paper), Gelman and Raghunathan (2001) and, briefly, in Little and Rubin (2002).

Note further that, as implemented in `stata` it is inappropriate for hierarchical data; generally if data are hierarchical, so should the imputation be. See the article by Carpenter and Goldstein for the multilevel modelling newsletter, downloadable from the preprints page on this site. More generally, we think the general application of this approach to hierarchical data is problematic.

Summary and conclusions

- Untestable assumptions unavoidable with missing data.
- Shun unprincipled methods.
- MI is most convenient under MAR.
  -To increase the chance that this is approximately true, we may wish to include several predictors of missingness that we do not want to adjust for in the final analysis.
- Multiple imputation is particularly useful for missing covarites, especially in:
  -survey settings where there is a separate imputer and analyst
  - large and messy problems, where a full likelihood or Bayesian analysis is impractical.

Summary and conclusions

- For models with missing responses, provided the covariates predictive of dropout are included, similar results are obtained to regression models (or mixed models, for longitudinal data).
  - in most missing outcome situations, preferable *not* to use multiple imputation, as it wastes information.
- Ideally, should consider a form of sensitivity analysis, though this is often not straightforward. - proper MI analyses are awkward under MNAR; it is necessary to make proper imputations from the posterior *conditional on the missing value indicator*.
  - Instead we can modify the imputation model to assess sensitivity, for example by using a postulated accept-reject mechanism on imputations.
- Often, serious thought unavoidable!