MULTIPLE IMPUTATION WITH SURVEY WEIGHTS: A MULTILEVEL APPROACH

M. QUARTAGNO*
J. R. CARPENTER
H. GOLDSTEIN

Multiple imputation is now well established as a practical and flexible method for analyzing partially observed data, particularly under the missing at random assumption. However, when the substantive model is a weighted analysis, there is concern about the empirical performance of Rubin's rules and also about how to appropriately incorporate possible interaction between the weights and the distribution of the study variables. One approach that has been suggested is to include the weights in the imputation model, potentially also allowing for interactions with the other variables. We show that the theoretical criterion justifying this approach can be approximately satisfied if we stratify the weights to define level-two units in our data set and include random intercepts in the imputation model. Further, if we let the covariance matrix of the variables have a random distribution across the level-two units, we also allow imputation to reflect any interaction between weight strata and the distribution of the variables. We evaluate our proposal in a number of simulation scenarios, showing it has promising performance both in terms of coverage levels of the model parameters and bias of the associated Rubin's variance estimates. We illustrate its application to a weighted analysis of factors predicting reception-year readiness in children in the UK Millennium Cohort Study.

M. Quartagno, J. R. Carpenter, and H. Goldstein are with Department of Medical Statistics, London School of Hygiene & Tropical Medicine, Keppel Street, London. M. Quartagno and J. R. Carpenter are with the MRC Clinical Trials Unit at UCL, 90 High Holborn, London. H. Goldstein is with Center for Multilevel Modelling, University of Bristol, 35 Berkeley Square, Bristol. Matteo Quartagno was supported by funding from the European Community's Seventh Framework Programme FP7/2011: Marie Curie Initial Training Network MEDIASRES ("Novel Statistical Methodology for Diagnostic/Prognostic and Therapeutic Studies and Systematic Reviews," mediasres-itn.eu) with the Grant Agreement Number 290025. James Carpenter is funded by the MRC Hub for Trials Methodology and supported by MC_UU_12023/21 and MC_UU_12023/29. This article was completed during a research visit to the Malawi Epidemiology and Research Unit.
*Address correspondence to M. Quartagno, MRC Clinical Trials Unit at UCL, 90 High Holborn, WC1V 6LJ, London, UK; E-mail: m.quartagno@ucl.ac.uk.

doi: 10.1093/jssam/smz036

KEYWORDS: Missing data; Multilevel; Multiple imputation; Survey; Weights.

1. INTRODUCTION

When collecting data for research, it is often the case that we are not able to obtain all the desired information for various reasons (e.g., lack of resources, unwillingness to disclose information, loss to follow-up). Unfortunately, such missing data complicate the intended analysis, not only causing a loss of power but also potentially biasing the results—when the reason for the missing data is associated with our scientific question.

For this reason, many methods have been developed to deal with missing data, each relying on a series of different assumptions. One of the biggest categories of missing data methods is represented by imputation strategies. Imputing missing data means replacing the missing values with a particular value, drawn from a specified distribution, typically from the conditional distribution of the missing data given the observed data. Fitting the substantive analysis model to such an imputed dataset gives the same weight to observed and imputed values; however, the latter are, at best, good guesses, and therefore, they should be somehow down-weighted. Otherwise, such an approach will result in marked underestimation of the standard errors because of a failure to reflect uncertainty due to the missing values.

In some specific settings, methods to obtain a valid variance estimate under single imputation have been derived (Rao and Shao 1992; Särndal, Swensson, and Wretman 1992; Rao 1996; Beaumont, Haziza, and Bocci 2011), and these are often used to handle missing data in surveys.

Alternatively, a very flexible, general method to address the same issue is multiple imputation (MI) Rubin (1987). With MI, the missing values are imputed from the Bayesian predictive distribution of the missing data, given the observed data, to create K imputed datasets. The substantive analysis model is then fitted to each of these in turn, giving K different estimates of the model parameters $\hat{\theta}_k$ together with their standard error estimates $\hat{\sigma}_k$. These are combined for final inference using Rubin's rules:

$$\hat{\theta}_{MI} = \sum_{k=1}^{K} \hat{\theta}_k / K,$$

$$Var(\hat{\theta}_{MI}) = \frac{1}{K}\sum_{k=1}^{K} \hat{\sigma}_k^2 + \left(1 + \frac{1}{K}\right) \sum_{k=1}^{K} (\hat{\theta}_k - \hat{\theta}_{MI})^2.$$

Over the last 25 years, the practicality and flexibility of MI, coupled with the availability of accessible software, has led it to become increasingly popular, particularly in clinical research.

As with all statistical methods, MI relies on some assumptions; in particular, most MI methods assume data are Missing at Random (MAR) (Rubin 1976), which broadly means that the reason for their missingness is unrelated to the unseen values, after conditioning on all the observed data. Another important issue is that of congeniality. This was raised in the original work from Rubin and was thoroughly investigated in a series of articles in the mid-nineties (Fay 1992, 1993; Meng 1994). These highlighted that in order for MI to lead to valid inference, the imputation model (i.e., the model used to impute the data) and the substantive model (i.e., the original model we wanted to fit on the complete data) must be congenial, which means, loosely speaking, that they need to be derived from the same joint model. In some situations, particularly when the substantive model has nonlinear effects or interactions, it can be challenging to choose a congenial imputation model (Goldstein, Carpenter, and Browne 2014; Bartlett, Seaman, White, and Carpenter 2015; for a practical review of the issue, see Carpenter and Kenward 2013, pp. 64–73).

Another issue that was discussed in the study by Meng (1994) is that of self-efficiency; a procedure is self-efficient if it is not possible to gain precision by applying it to a subset of the whole data. Self-efficiency of the complete-data procedure is required for the validity of MI inference (Meng and Romero 2003).

This article focuses on the situation in which the substantive model is a weighted regression model; this is common in survey sampling settings, where appropriate weighting is often used to take account of the sampling schemes (e.g., Särndal et al. 1992). Throughout, we assume the weights considered are the final ones, after adjustments for nonresponse (Holt and Elliot 1991) and calibration (Deville and Särndal 1992). The idea of MI was originally expounded in a survey setting, and in Rubin (1987), it was implicitly assumed that the imputer should have access to the variables used to construct any weights and should always include them in the imputation model. However, discussion of weighting was limited to a brief reference in the introduction, where it was noted that "[weighting's] apparent simplicity disappears with multivariate outcomes", followed by two excercises. Two questions remained unanswered at the time:

(i) How should we include weights in the imputation model?
(ii) Does Rubin's variance formula still hold in these settings?

To answer these questions, it is important to clarify the inferential framework under which properties of MI are to be evaluated. For example, if evaluating the properties with respect to the joint distribution of the response mechanism and the sampling mechanism, Rubin's variance estimator is valid under the assumption of proper imputation (Rubin 1987). Kim, Michael Brick Fuller, and Kalton (2006) took a different approach and evaluated the properties with respect to the joint distribution of the sampling mechanism, the response mechanism, the imputation mechanism, and a super-population model. They showed that, because of a lack of self-efficiency, even in the case of a simple weighted regression model with missing data in the outcome only, using a standard imputation model with Rubin's rules results in an upwardly biased estimate of the variance. In particular, assuming both our substantive and imputation model are of the form:

$$Y \sim N(\mathbf{X}\boldsymbol{\theta}, \mathbf{I}\sigma^2),$$

and that we want to use the weighted least square estimator:

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y},$$

then the bias in the Rubin's variance estimator is

$$Bias(V_{MI}) = 2\sigma^2\left(\mathbf{w}_m^T\mathbf{w}_m - tr(\mathbf{w}_o^T\mathbf{X}_o(\mathbf{X}_o^T\mathbf{X}_o)^{-1}\mathbf{X}_m^T\mathbf{w}_m)\right),$$

where σ^2 is the residual variance, the subscript "m" indexes missing observations, and "o" to the observed, so that for example, $\mathbf{X}_o$ represents the set of covariates X for complete records. As Kim et al. (2006) pointed out, a practically important consequence follows from this expression: the bias vanishes if the weights are included in the space spanned by the variates in the regression model (i.e., for $\mathbf{w}_o = \mathbf{X}_o\,\mathbf{d}$ and $\mathbf{w}_m = \mathbf{X}_m\,\mathbf{d}$, for some value of $\mathbf{d}$). Therefore, a simple way to correct for the bias in Rubin's variance estimator when using survey weights, at least with linear regression, is to introduce these into the linear predictor.

However, Kim et al. (2006) also picked up the result of Meng (1994) that the MI variance estimator is biased upward if the imputer assumes more than the analyst; therefore, for accurate inference within domains for survey data, the imputer needs to include in his model (i) the weights, (ii) all the domain indicators (i.e., all relevant covariates of the weighted regression), and (iii) their interactions for valid MI inference in general.

Seaman et al. (2012) extended Kim's results, showing that if using multiple imputation with this correctly specified model, Rubin's variance estimate is asymptotically unbiased with missing data in the outcome only. With missing covariates, there is generally an upward bias in the variance; however, the simulation results they report suggest this is of little practical concern.

In practice, a key issue is the correct specification of the imputation model, taking into account the weights and their interaction with the covariates. Ideally, we would impute separately in each weight strata so that the relationship between the variables are allowed to differ across weight strata. Thus ideally, we should include weight as a factor variable, possibly with interactions with other variables, alongside including it as a linear variable to avoid the bias in Rubin's MI variance estimator. In general, this will make estimation noisy because of the increased number of parameters.

Instead in this article, we propose using the weight variable to define a second level and then adopt a multilevel approach.

In section 2, we describe our proposed approach and how it approximately satisfies Kim et al.'s (2006) criteria. Then in section 3, we describe a set of simulation studies to evaluate our proposal against alternative approaches. The results are given in section 4. In section 5, we apply the same methods to handle missing items in wave 2 of the Millennium Cohort Study; finally, we conclude in section 6 with a discussion.

2. METHODS

As set out in the previous section, the situation we are interested in is a sample survey dataset, provided with weights and affected by missing items. This is exemplified by the Millennium Cohort Study we analyze in section 5. Our setup follows Kim et al. (2006); we assume we have a complex sample from an infinite super-population. We evaluate the properties of Rubin's MI variance estimator considering the joint distribution of the sampling, response, imputation, and super-population models.

2.1 Proposed Approach

We have already noted that the bias in Rubin's MI variance formula, given by (5), vanishes if the weights are included in the space of the variates spanned in the regression model (i.e., if $\mathbf{w}_o = \mathbf{X}_o\mathbf{d}$ and $\mathbf{w}_m = \mathbf{X}_m\,\mathbf{d}$, for some value of $\mathbf{d}$).

For example, suppose that we have p covariates for each unit, of which the first is the intercept and the second the weight. Then if the $p \times 1$ vector $\mathbf{d} = (0, 1, 0, \ldots, 0)^T$, this criterion is satisfied. However, while this is sufficient for valid variance estimation using Rubin's rules for a mean, as noted previously, it is insufficient when we have domains in our data; then we need to include both the domain indicators and their interactions with the weights.

Instead, now suppose that we group the weights, without loss of generality, into $g = 1, \ldots, G$ groups. We include additional G dummy variables as the leftmost covariates in $\mathbf{X}$ indexing which of the groups unit i's weight belongs to. Also, let $\mathbf{d} = (\bar{w}_1, \bar{w}_2, \ldots, \bar{w}_G, 0, \ldots, 0)^T$, where $\bar{w}_g$ is the mean weight for group g. Now the criteria for the bias in the variance vanishing is approximately satisfied. Further, the approximation will improve as the weight SD within the groups decreases. This is often possible to do in applications because the weights are calculated (often by the data provider) from a set of categorical predictors. Importantly, we also note that including these extra parameters in the regression model of interest changes from one that is marginal to the weights to one that conditions on them and that this may not be the model of concern for the analyst.

While approximately satisfying the criterion for Rubin's variance formula to work, this approach also has the advantage that it does not require the relationship between the weights and the dependent variable to be linear; it is unstructured across the groups. However in general, fitting a large number of fixed parameters is not desirable nor is it consistent with the aims of the data analyst, as pointed out previously.

Instead, we propose letting the G weight groups define a second level in the data and including random intercepts. This still approximately satisfies the criterion for bias in (5) to vanish, but now we can pool information across weight groups where appropriate. In other words, when we impose the standard assumption that the random intercept distribution has zero mean, the fixed part of the model will represent the (marginal) expected relationship for the population, as desired by the analyst.

This is not sufficient in general, though, because ideally (as noted in the introduction) we should allow for an interaction between the weights and the other variables in the imputation model. We can do this by allowing the covariance matrix of the (level-one) variables to vary across the (level-two) units (weight-strata). Again, rather than introduce a lot of parameters, we can give the covariance matrix a random distribution across strata.

For a specific example of our proposal, suppose that the substantive model is a weighted linear regression of $y_{i,j}$ on $x_{1,i,j}, x_{2,i,j}$, where $i = 1, \ldots, n_j$ indexes units in strata $j = 1, \ldots, J$ with weight w_j. Suppose data are MAR. We let the weight strata define level two, and our imputation model is:

$$\begin{gathered}
\begin{pmatrix} y_{i,j} \\ x_{1,i,j} \\ x_{2,i,j} \end{pmatrix} \sim N\left(\begin{pmatrix} \theta_{0,0,j} + u_{0,j} \\ \theta_{1,0,j} + u_{1,j} \\ \theta_{2,0,j} + u_{2,j} \end{pmatrix}, \mathbf{\Omega}_j \right) \\
\begin{pmatrix} u_{0,j} \\ u_{1,j} \\ u_{2,j} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \mathbf{\Psi} \right) \qquad , \\
\mathbf{\Omega}_j \sim W^{-1}(a, A^{-1}),
\end{gathered}$$

where W^{-1} denotes the inverse Wishart distribution. Notice this includes the random intercepts $u_{k,j}$ for each variable k and that the level-1 covariance matrix

Ω_j varies across weight strata j, allowing the association of y, x_1, x_2 to vary across strata. $\theta_{k,0,j}$ represent the overall means of the three variables, while Ψ is the level-two covariance matrix. This model was proposed in different context by Yucel (2011) and developed for individual patient data meta-analysis by Quartagno and Carpenter (2016).

Compared with including the weights as a linear term in the imputation model, together with their interaction with the other variables, model (6) has the advantage that the relationship across the weight strata is not required to be linear; it is driven by the data, and information is pooled across strata as appropriate. While in general it only approximately satisfies the criteria for Rubin's variance formula to hold, we will show by simulations that the difference between the empirical and Rubin's MI variance is small or negligible, and suggesting this will be satisfactory in applications.

Note that if the weight is common in each group G, then as the number of observations in each strata gets large, this approach tends to the natural—and often optimal—approach of imputing separately in each strata. However, if the proportion of missing in some strata is high, our approach may be able to improve on this.

Having outlined our proposal, we now evaluate it using a series of simulation studies, comparing with imputing separately in each strata, ignoring the weights in the imputation, and including them in various ways.

3. SIMULATION STUDIES

First, we describe the base-case simulation scenario, before outlining the methods we are comparing with (6), and briefly discuss their relative merits. We conclude this section by describing three additional simulation scenarios.

The simulation scenarios are designed to reveal differences between the methods. In all the scenarios, we consider three variables: Y, X_1, and X_2; our total sample size is 400 individuals, stratified in ten equal-sized strata, each with corresponding known weight.

3.1 Base-Case Scenario

The base-case scenario simulated data from the following mechanism:

$$\begin{pmatrix} x_{1,i,j} \\ x_{2,i,j} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0.5 & 0.2 \\ 0.2 & 0.5 \end{pmatrix} \right)$$

$$e_{i,j} \sim N(0,1)$$

$$\boldsymbol{\beta}_0 = (1,2,3,4,5,6,7,8,9,10)$$

$$\boldsymbol{\beta}_1 = 0.2\boldsymbol{\beta}_0$$

$$\boldsymbol{\beta}_2 = 0.5\boldsymbol{\beta}_0$$

$$y_{i,j} = \beta_{0,j} + \beta_{1,j}x_{1,i,j} + \beta_{2,j}x_{2,i,j} + e_{i,j}$$

Here, j indexes different weight strata. After having generated the data, the substantive analysis model is a weighted linear regression, where Y is the dependent variable, X_1 and X_2 are the covariates, and we assume the weights are known and equal to:

$$w = \left(\frac{1}{0.1}, \frac{1}{0.2}, \ldots, \frac{1}{1} \right) = (10, 5, \ldots, 1).$$

This may seem an extreme choice and in applications weights (and fixed effect parameters $\boldsymbol{\beta}$) would probably be more homogenous; however, we decided to use such extreme values in order to bring out the properties of the methods.

We simulate 1,000 data sets and make Y Missing Completely at Random (MCAR) with probability 0.5. We compare analysis of the original full data (FD) and complete records (CR) and use the competing multiple imputation methods we now describe. All of the imputation models are fitted by means of a Gibbs sampling algorithm using data augmentation to impute the missing data, using the R-package jomo (Quartagno, Grund, and Carpenter 2018).

3.2 Imputation Methods

We now describe the seven imputation approaches that we compare.

3.2.1 Multiple imputation with no weights (MI-noW). Multiple imputation with no weights (MI-noW) uses the first and simplest imputation model we might consider. It consists of a multivariate normal model for the three partially observed variables and does not make any use of the weights:

$$\begin{pmatrix} y_{i,j} \\ x_{1,i,j} \\ x_{2,i,j} \end{pmatrix} \sim N(\boldsymbol{\theta}, \boldsymbol{\Omega}).$$

We know from section 1 that weights should be included in the imputation model for Rubin's rules to hold, and therefore, we expect this method to perform relatively poorly.

3.2.2 Multiple imputation with weights (MI-W). The next option is to use an imputation model where the weights are included as additional variables; the easiest way to do this is to include them as an additional covariate in the multivariate normal imputation model, assuming a linear relation between weights and all three variables:

$$\begin{pmatrix} y_{i,j} \\ x_{1,i,j} \\ x_{2,i,j} \end{pmatrix} \sim N\left(\begin{pmatrix} \theta_{0,0,j} + \theta_{0,1,j} w_j \\ \theta_{1,0,j} + \theta_{1,1,j} w_j \\ \theta_{2,0,j} + \theta_{2,1,j} w_j \end{pmatrix}, \boldsymbol{\Omega}\right).$$

While this model includes the weights it (i) assumes a linear relationship between the weights and the variables and (ii) does not include the interactions between the weights and the covariates that appear in the substantive model, which according to the literature is desirable, as seen in the introduction.

3.2.3 Multiple imputation with weights and interactions (MI-xW). As outlined in the introduction, the literature (Kim et al. 2006; Seaman et al. 2012) suggests a better imputation model should include not only the weights but also all interactions between weights and covariates. This can be done easily when missing data are confined to the outcome variable—but not when data are missing in all variables. In this setting, we need to use the substantive model compatible imputation developed by Goldstein et al. (2014) (see also Bartlett et al. 2015).

The idea is to use an imputation model that partitions the joint distribution of the three variables between a joint distribution for the covariates and a conditional distribution of the dependent variable given the covariates:

$$\begin{pmatrix} x_{1,i,j} \\ x_{2,i,j} \end{pmatrix} \sim N\left(\begin{pmatrix} \theta_{1,0,j} + \theta_{1,1,j} w_j \\ \theta_{2,0,j} + \theta_{2,1,j} w_j \end{pmatrix}, \boldsymbol{\Omega}\right) .$$

$$y_{i,j} | x_{1,i,j}, x_{2,i,j} = \beta_0 + \beta_1 x_{1,i,j} + \beta_2 x_{2,i,j} + \beta_3 w_j + \beta_4 w_j x_{1,i,j} + \beta_5 w_j x_{2,i,j} + \epsilon_{i,j}$$

$$\epsilon_{i,j} \sim N(0, \sigma^2)$$

Missing data in Y are imputed from the conditional model given the covariates, weights, and their interactions, while missing data in the covariates are imputed compatibly with the model for Y, by means of a Metropolis-Hastings step within the Gibbs sampler. Note the model specified for Y in

(11) is not the substantive model, which is the weighted regression of y on x_1, x_2.

Although this approach should improve on (1) and (2), there are two potential shortcomings. First, when values are missing in all three variables, we do not have an interaction with the weights and the distribution of $x_2|x_1$ and vice versa because their covariance matrix, Ω, is common across the strata. Second, it again assumes a linear relationship of y on the weights and their interactions; with a large number of covariates, the conditional model for Y becomes complicated, and estimating the parameters may lead to noisy results.

3.2.4 Stratum-specific multiple imputation (MI-S). Where we have well-defined strata and sufficient data in each, this is perhaps the best approach. It is straight forward (we use standard imputation in each strata), allows a full interaction in the relationship between the variables by strata, and satisfies the criteria for Rubin's rules to give valid inference. For the ten strata in our simulated data, we therefore have,

$$\begin{pmatrix} y_{i,j} \\ x_{1,i,j} \\ x_{2,i,j} \end{pmatrix} \sim N\left(\boldsymbol{\theta}_{\mathrm{j}}, \boldsymbol{\Omega}_j\right) \qquad j = 1, \ldots, 10.$$

The disadvantage of this method is that it may struggle with small strata or substantial numbers of missing values within some strata.

3.2.5 Homoscedastic multilevel multiple imputation (MLMI-Hom). This is the first of our three multilevel imputation approaches. This approach does not use the random weight-strata covariance matrices in (6); the reason for this is to explore if this aspect is necessary. Thus, the imputation model is (6) with common covariance matrix:

$$\begin{pmatrix} y_{i,j} \\ x_{1,i,j} \\ x_{2,i,j} \end{pmatrix} \sim N\left(\begin{pmatrix} \theta_{0,0,j} + u_{0,j} \\ \theta_{1,0,j} + u_{1,j} \\ \theta_{2,0,j} + u_{2,j} \end{pmatrix}, \boldsymbol{\Omega}\right)$$

$$\begin{pmatrix} u_{0,j} \\ u_{1,j} \\ u_{2,j} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \boldsymbol{\Psi}\right),$$

where the weight strata form the $j = 1, \ldots, J$ level-two groups. The problem with this method is that the level-one correlation between the outcome and the covariates of the substantive model is kept fixed in the imputation model across strata; this may bias inferences when (as will often be the case in practice) the association between variables varies with the weights.

3.2.6 Heteroscedastic multilevel multiple imputation (MLMI-Het). The sixth approach is (6), discussed in section 2. While not necessarily optimal in all scenarios, it should have good performance across them all. In particular, it allows the relationships between the variables to vary with the weights but does not insist this happens in a linear way.

3.2.7 Substantive model compatible multilevel MI (MLMI-SMC). The seventh and final method is to use multilevel substantive model compatible imputation (Goldstein et al. 2014). Essentially, this makes method three, (11), multilevel by giving the coefficients in the model of $y|x_1, x_2$ random coefficients across the weight strata:

$$\begin{pmatrix} x_{1,i,j} \\ x_{2,i,j} \end{pmatrix} \sim N\left(\begin{pmatrix} \theta_{1,0,j} + u_{1,j} \\ \theta_{2,0,j} + u_{2,j} \end{pmatrix}, \mathbf{\Omega} \right)$$

$$\begin{pmatrix} u_{1,j} \\ u_{2,j} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \mathbf{\Psi} \right)$$

$$y_{i,j}|x_{1,i,j}, x_{2,i,j} = \beta_0 + v_{0,j} + (\beta_1 + v_{1,j})x_{1,i,j} + (\beta_2 + v_{2,j})x_{2,i,j} + \epsilon_{i,j}\,,$$

$$\epsilon_{i,j} \sim N(0, \sigma^2)$$

$$\begin{pmatrix} v_{0,j} \\ v_{1,j} \\ v_{2,j} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \mathbf{\Psi}_Y \right)$$

where $v_{k,j}$ are independent from $u_{k,j}$. This model is almost as flexible as (6), but not quite, as the distribution of (x_1, x_2) does not fully vary across weight strata.

3.3 Further Simulation Scenarios

In addition to the base-case scenario with missing values in Y alone described at the start of this section, we consider four further cases:

3.3.1 Base-case scenario: missingness in Y, X_1, X_2. We use the base-case scenario (7), but this time with 20 percent missing data in all three variables to have again approximately 50 percent complete records (0.8 * 0.8 * $0.8 = 0.512 = 51.2$ percent).

3.3.2 Base-case scenario: missingness proportional to weights. We use the base-case data-generating model, but now with the proportion of missing data proportional to the weights:

$$p_{miss,s} = \frac{2w_s}{\sum_{s=1}^{10} w_s}.$$

3.3.3 Base-case scenario: missingness inversely proportional to weights. We use the base-case data-generating model, but now with the proportion of missing data *inversely* proportional to the weights:

$$p_{miss,s} = \frac{\frac{2}{w_s}}{\sum_{s=1}^{10} \frac{1}{w_s}}.$$

3.3.4 GLM scenario. Here, instead of a continuous dependent variable y, we simulate a binary dependent variable as follows:

$$\begin{gathered}
\begin{pmatrix} x_{1,i,j} \\ x_{2,i,j} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0.5 & 0.2 \\ 0.2 & 0.5 \end{pmatrix} \right) \\
\boldsymbol{\beta}_0 = 0.1 \times (1,2,3,4,5,6,7,8,9,10) \\
\boldsymbol{\beta}_1 = -\boldsymbol{\beta}_0 \qquad\qquad\qquad . \\
\boldsymbol{\beta}_2 = 0.5\boldsymbol{\beta}_0 \\
\pi_{i,j} = \Pr(Y_{i,j} = 1) = \frac{1}{1 + \exp\{-(\beta_{0,j} + \beta_{1,j}x_{1,i,j} + \beta_{2,j}x_{2,i,j})\}} \\
y_{i,j} \sim \text{Bernoulli}(\pi_{i,j}).
\end{gathered}$$

Under data-generating model (15) applying the same weights as before, the true parameter values are

$$\begin{gathered}
\beta_0 = 0.334 \\
\beta_1 = -0.334 \\
\beta_2 = 0.170.
\end{gathered}$$

Data are made MCAR independently in each variable with probability 0.2. When imputing data generated using this model, all the methods apply the latent normal method for imputing the binary dependent variable (Carpenter and Kenward 2013, Chapter 5; Quartagno and Carpenter 2019).

3.3.5 Realistic scenario. As discussed, the base-case simulation scenario involved the choice of very extreme simulation parameters to make the comparison of the performance of different methods clear. In this last scenario, we instead use more realistic parameters, mimicking the distribution of data from the Millennium Cohort Study. In particular, we generate data for a total of 5,400 individuals divided in nine strata; weights associated with each stratum range between 0.23 and 2, and we use the following data-generating mechanism:

$$\begin{pmatrix} x_{1,i,j} \\ x_{2,i,j} \end{pmatrix} \sim N\left(\begin{pmatrix} 2.95 \\ 0.94 \end{pmatrix}, \begin{pmatrix} 0.76 & -0.01 \\ 0.01 & 1.11 \end{pmatrix}\right)$$

$$e_{i,j} \sim N(0, 0.91)$$

$$\boldsymbol{\beta}_{0,j} \sim N(-0.63, 0.3)$$

$$\boldsymbol{\beta}_{1,j} \sim N(0.30, 0.3)$$

$$\boldsymbol{\beta}_{2,j} \sim N(-0.21, 0.3)$$

$$y_{i,j} = \beta_{0,j} + \beta_{1,j}x_{1,i,j} + \beta_{2,j}x_{2,i,j} + e_{i,j}.$$

4. SIMULATION RESULTS

All the simulations used 1,000 replications. Imputation used the jomo package, generating twenty imputed tables, with a burn-in of 500, and 500 updates between each imputed dataset.

4.1 Base-Case Scenario

The results of the base-case scenario, with 50 percent of y values MCAR but other variables complete, are shown in the top part of table 1. As expected, MI-S performs best here, giving approximately unbiased point estimates and good coverage levels close to 95 percent. However, both multilevel imputation methods, MLMI-Het and MLMI-SMC, give similar results for bias, precision, and coverage. In particular, the model SE (i.e., the SE obtained using Rubin's rules) and the empirical SE are similar for all three parameter estimates. While MI-xW has similar results, there seems to be slightly more bias in the

Table 1. Simulation Results: Mean, Empirical SE, Model SE, and Coverage Level for the Three Fixed Effect Parameters of the Substantive Weighted Regression Model, β_0, β_1 and β_2

	β_0				β_1				β_2			
	Mean	Emp. SE	Mod. SE	Cov.	Mean	Emp. SE	Mod. SE	Cov.	Mean	Emp. SE	Mod. SE	Cov.
Missing in Y only												
True Value	3.414				0.685				1.707			
Full data	3.425	0.098	0.096	0.940	0.696	0.290	0.276	0.929	1.716	0.292	0.284	0.936
Complete records	3.438	0.139	0.206	0.994	0.697	0.409	0.384	0.925	1.719	0.405	0.401	0.931
MI-noW	4.465	0.138	0.298	0.001	0.895	0.438	0.473	0.948	2.220	0.437	0.480	0.830
MI-W	3.430	0.176	0.220	0.985	0.895	0.400	0.433	0.941	2.221	0.400	0.440	0.826
MI-xW	3.400	0.141	0.219	0.998	0.675	0.386	0.454	0.966	1.697	0.401	0.459	0.971
MI-S	3.423	0.121	0.128	0.961	0.688	0.306	0.308	0.947	1.623	0.307	0.313	0.946
MLMI-Hom	3.447	0.174	0.157	0.923	0.905	0.325	0.329	0.890	2.210	0.328	0.339	0.706
MLMI-Het	3.434	0.123	0.123	0.954	0.696	0.315	0.304	0.935	1.717	0.314	0.310	0.945
MLMI-SMC	3.431	0.121	0.120	0.954	0.695	0.310	0.300	0.940	1.720	0.313	0.306	0.938
Missing in all 3 variables												
True value	3.414				0.685				1.707			
Full data	3.425	0.097	0.096	0.944	0.702	0.284	0.276	0.936	1.705	0.288	0.284	0.948
Complete records	3.569	0.481	0.474	0.923	0.748	0.856	0.722	0.891	1.792	0.879	0.741	0.881

MI-noW	4.625	0.187	0.325	0.002	0.973	0.582	0.626	0.928	2.361	0.538	0.577	0.806
MI-W	3.456	0.208	0.259	0.984	0.917	0.517	0.571	0.954	2.128	0.532	0.589	0.915
MI-xW	3.503	0.221	0.281	0.991	0.668	0.472	0.563	0.977	1.680	0.480	0.586	0.990
MI-S	3.449	0.135	0.162	0.979	0.637	0.329	0.398	0.980	1.308	0.342	0.414	0.890
MLMI-Hom	3.484	0.213	0.199	0.921	0.978	0.422	0.431	0.892	2.281	0.429	0.428	0.732
MLMI-Het	3.455	0.150	0.152	0.950	0.685	0.420	0.407	0.929	1.629	0.420	0.417	0.943
MLMI-SMC	3.459	0.141	0.150	0.964	0.675	0.384	0.407	0.955	1.699	0.380	0.416	0.970

Note.—Data are generated from the base-case configuration and successively missing data are introduced either (i) in the outcome only (50 percent MCAR) or (ii) in all three variables 20 percent MCAR per each variable, leading to 50 percent complete records. We compare the true value of the parameters with the full data estimates and with the estimates handling missing data with CR, MI-noW, MI-W, MI-xW, MI-S, MLMI-Hom, MLMI-Het, and MLMI-SMC.

parameter estimates; also, the model SE is somewhat larger than empirical SE and greater than MI-S, MLMI-Het, and MLMI-SMC. Additionally, MI-noW, MI-W, and MLMI-Hom all have some shortcomings in the way they handle weights in the imputation model, and therefore, it is not surprising that they lead to unsatisfactory results.

Interestingly, the estimated SEs are consistently smaller with MI-S, MLMI-Het, and MLMI-SMC than for the CR analysis. In general, when missing data are in the outcome only, it is not possible to recover information without auxiliary variables. Here though, the combination of small strata with relatively high proportions of missing data is likely to be the reason why results after MI gain over CR.

4.2 Base-Case Scenario—Missingness in Y, $X1$, $X2$

The results are shown in the bottom half of table 1. The MIMI-Het and MIML-SMC now give the best results, outperforming MI-S because of their ability to pool information across strata through the random effects (this issue with MI-S also appears to affect CR and MI-xW, where the model SEs for β_1, β_2 are relatively large). The MIMI-Het and MIML-SMC also recover a nontrivial proportion of information compared with CR. Some of the parameter estimates with CR, MI-S, and MI-xW are now slightly more biased, possibly because of a common reason (i.e., the fact that strata are so small); this leads to undercovering for β_2 with CR and MI-S. Although with MI-xW, an overestimation of the standard error leads to overcovering, despite the bias introduced.

Finally, we note that in both these scenarios, the full flexibility of MLMI-Het is not needed, as the covariance matrix of the covariates is common across strata in the data-generating mechanism. However, this does not adversely affect its performance.

4.3 Base-Case Scenario—Missingness Proportional to Weights

This is a challenging scenario because we have a nontrivial proportion of missing data in each variable, and missingness is proportional to the weights that are inversely proportional to the strata coefficients. The top parts of table 2 and figure 1 summarize the results.

The best results are now obtained with MLMI-Het and MLMI-SMC, which both perform better than MI-S. This can be clearly seen in the top row of figure 1, which shows the three zip plots (Morris 2016) for MI-S, MLMI-Het, and MLMI-SMC. Each zip plot is for β_2 (true value 1.707) and ranks the results of the 1,000 replications top to the bottom according to their p value against the null hypothesis. The red vertical line indicates the true value, and we can see that both multilevel imputation methods seem approximately unbiased. The purple bars indicate simulations that failed to cover the true value in

Table 2. Simulation Results: Mean, Empirical SE, Model SE, and Coverage Level for the Three Fixed Effect Parameters of the Substantive Weighted Regression Model, β_0, β_1 and β_2

	β_0				β_1				β_2			
	Mean	Emp. SE	Mod. SE	Cov.	Mean	Emp. SE	Mod. SE	Cov.	Mean	Emp. SE	Mod. SE	Cov.
Missingness directly proportional to weights												
True value	3.414				0.685				1.707			
Full data	3.428	0.097	0.096	0.946	0.682	0.278	0.276	0.942	1.723	0.292	0.283	0.939
Complete records	5.343	0.177	0.199	0.000	1.077	0.372	0.355	0.775	2.676	0.405	0.372	0.302
MI-noW	5.182	0.133	0.272	0.000	1.204	0.341	0.448	0.846	2.775	0.341	0.432	0.285
MI-W	2.435	0.247	0.316	0.070	1.102	0.580	0.619	0.907	1.899	0.653	0.726	0.961
MI-xW	2.142	0.478	0.509	0.281	0.100	0.763	0.801	0.926	0.771	0.667	0.923	0.939
MI-S	3.476	0.155	0.182	0.974	0.796	0.353	0.422	0.961	1.461	0.420	0.501	0.958
MLMI-Hom	3.567	0.283	0.248	0.859	1.187	0.335	0.355	0.729	2.650	0.397	0.394	0.353
MLMI-Het	3.527	0.175	0.191	0.928	0.781	0.450	0.445	0.924	1.656	0.483	0.482	0.942
MLMI-SMC	3.535	0.157	0.193	0.959	0.717	0.354	0.393	0.958	1.783	0.393	0.432	0.941
Missingness inversely proportional to weights												
True value	3.414				0.685				1.707			
Full data	3.427	0.098	0.095	0.941	0.693	0.291	0.277	0.937	1.721	0.278	0.285	0.947
Complete records	2.661	0.116	0.127	0.000	0.536	0.291	0.272	0.894	1.329	0.296	0.279	0.706

Continued

Table 2. *Continued*

	β_0				β_1				β_2			
	Mean	Emp. SE	Mod. SE	Cov.	Mean	Emp. SE	Mod. SE	Cov.	Mean	Emp. SE	Mod. SE	Cov.
MI-noW	3.336	0.101	0.132	0.961	0.664	0.329	0.308	0.924	1.668	0.319	0.308	0.941
MI-W	3.395	0.099	0.113	0.972	0.661	0.314	0.304	0.940	1.681	0.309	0.305	0.945
MI-xW	3.391	0.102	0.114	0.965	0.702	0.321	0.301	0.928	1.741	0.308	0.302	0.943
MI-S	3.430	0.098	0.101	0.958	0.655	0.301	0.293	0.941	1.667	0.298	0.295	0.943
MLMI-Hom	3.427	0.100	0.101	0.955	0.656	0.313	0.295	0.926	1.654	0.306	0.295	0.938
MLMI-Het	3.427	0.100	0.100	0.951	0.681	0.304	0.290	0.935	1.707	0.302	0.293	0.941
MLMI-SMC	3.428	0.100	0.100	0.956	0.688	0.303	0.290	0.925	1.705	0.300	0.292	0.945

Note.—Data are generated from the base-case configuration and successively missing data are introduced either (i) with missingness directly proportional to weights or (ii) inversely proportional. We compare the true value of the parameters with the full data estimates and with the estimates handling missing data with CR, MI-noW, MI-W, MI-xW, MI-S, MLMI-Hom, MLMI-Het, and MLMI-SMC.

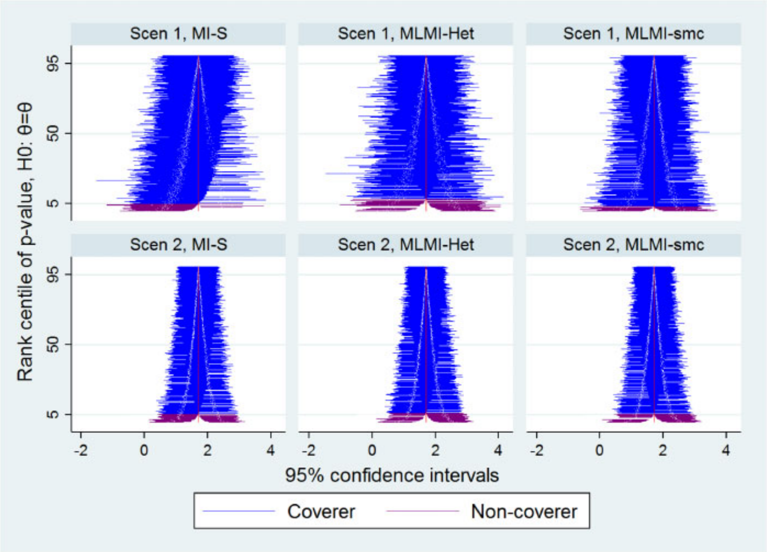

Figure 1. Zip Plot Comparing Simulation Results for MI-S, MLMI-Het, and MLMI-SMC for β_2, True Value 1.707. Top row: base-case scenario with missingness probability proportional to weights; bottom row: base-case scenario with missingness inversely proportional to weights.

their interval. These are approximately 5 percent for both multilevel methods; however, MLMI-Het tend to cover a bit less than 95 percent and MLMI-SMC a bit more. As expected, the fact that MLMI-Het allows for different covariate covariance across the strata (which is not present here) gives slightly larger SEs (wider CIs) than MIMI-SMC.

By contrast, the zip plot for MI-S has approximately correct coverage levels, but the mean estimate is slightly biased, as can be seen from the fact that the noncovering simulations are almost all to the left of the true value. This is likely due to the fact that in this scenario, the probability of missingness is extremely high for the most weighted strata, which are the ones with the biggest effect on the overall weighted estimate of the parameters. When imputing using MI-S, we therefore do not have enough information in some strata to build and fit our stratum-specific imputation model, leading to biased estimates, mainly toward the null.

Because missingness is no longer MCAR here, CR is not valid. Also (table 2, top half), MI-xW gives poor results here; the pattern of missing data means it leans heavily on its incorrect assumption of a linear effect of the weights.

4.4 Base-Case Scenario—Missingness Proportional to Weights

This scenario is less challenging than the previous one because the proportion of missing data is higher in the strata with lower weight. The results are shown in the lower parts of table 2 and figure 1. Once again, we see good results for MIMI-Het and MLMI-SMC; however, relative to these, the performance of MI-S and MI-xW has improved.

4.5 GLM Scenario

Recall that in this scenario, once again each variable is MCAR with probability 0.2. Table 3 shows the results. There is a little bias in all the coefficient estimates, most likely due to small sample effects in GLMs. Here, MI-xW, MI-S, MLMI-Het, and MLMI-SMC are all competitive, with best results for MI-xW and MLMI-SMC. However, for MI-xW the model SEs tend to be smaller than the empirical SEs; this is avoided with MLMI-SMC and MLMI-Het. For MI-S, the model SEs are also larger than the empirical SEs, and this allows the coverage to be relatively good despite the slight bias (particularly for β_2).

4.6 Realistic Scenario

Results are again shown in table 3. While generally inference seems to be acceptable with most imputation methods, as indicated by negligible biases and good coverage levels, MI-xW, MLMI-Het, and MLMI-SMC are the best methods for variance estimation, as they are the methods for which model and empirical standard errors are most similar. MI-S seems to work similarly well, as expected given that weight strata are large in this example (i.e., 600 observations per stratum), and hence, within-stratum imputation is not as noisy as in the previous examples.

4.7 Summary of Simulation Results

In summary, both the MLMI-Het and MLMI-SMC give similar results across the range of scenarios considered here, and in each scenario, either are competitive with the best method or give the best results. In particular, neither MI-S nor MI-xW give such consistently good results.

5. APPLICATION TO MILLENNIUM COHORT DATA

We now use the methods evaluated previously in an analysis of the Millennium Cohort Study dataset (Plewis 2007). This is a multidisciplinary research project following the lives of around 19,000 children born in the UK in 2000 and 2001. We focus on the second wave (children around three years of age), where some items are missing, particularly in the family income and hearing problems variables (around 12 percent missing).

Table 3. Simulation Results for the GLM and Realistic Scenarios: Mean, Empirical SE, Model SE, and Coverage Level for the Three Fixed Effect Parameters of the Substantive Weighted Regression Model, β_0, β_1 and β_2

	β_0				β_1				β_2			
	Mean	Emp. SE	Mod. SE	Cov.	Mean	Emp. SE	Mod. SE	Cov.	Mean	Emp. SE	Mod. SE	Cov.
GLM scenario												
True value	0.334				−0.334				0.170			
Full data	0.331	0.146	0.139	0.943	−0.337	0.222	0.219	0.943	0.164	0.226	0.216	0.933
Complete records	0.348	0.202	0.196	0.948	−0.333	0.309	0.307	0.943	0.163	0.309	0.302	0.947
MI-noW	0.379	0.149	0.157	0.958	−0.409	0.240	0.274	0.979	0.207	0.238	0.271	0.977
MI-W	0.341	0.162	0.158	0.937	−0.406	0.239	0.274	0.980	0.204	0.236	0.271	0.982
MI-xW	0.340	0.162	0.157	0.943	−0.334	0.284	0.281	0.949	0.161	0.285	0.278	0.947
MI-S	0.339	0.162	0.156	0.939	−0.288	0.249	0.267	0.955	0.131	0.249	0.263	0.963
MLMI-Hom	0.348	0.161	0.157	0.942	−0.399	0.240	0.272	0.984	0.200	0.234	0.270	0.979
MLMI-Het	0.349	0.159	0.157	0.947	−0.366	0.258	0.274	0.971	0.180	0.256	0.270	0.963
MLMI-SMC	0.341	0.162	0.158	0.941	−0.358	0.270	0.279	0.956	0.177	0.270	0.276	0.962
Realistic scenario												
True value	−0.821				0.360				−0.460			
Full data	−0.821	0.056	0.056	0.952	0.360	0.017	0.018	0.951	−0.460	0.015	0.014	0.944
Complete records	−0.821	0.061	0.062	0.942	0.360	0.019	0.020	0.953	−0.460	0.017	0.016	0.946
MI-noW	−0.821	0.058	0.062	0.956	0.360	0.018	0.020	0.967	−0.460	0.016	0.016	0.957
MI-W	−0.821	0.058	0.062	0.958	0.360	0.018	0.020	0.965	−0.460	0.016	0.016	0.957
MI-xW	−0.821	0.062	0.063	0.950	0.360	0.019	0.020	0.956	−0.460	0.017	0.016	0.950
MI-S	−0.821	0.062	0.063	0.945	0.360	0.019	0.020	0.952	−0.460	0.017	0.016	0.946
MLMI-Hom	−0.822	0.058	0.062	0.958	0.360	0.018	0.020	0.960	−0.460	0.016	0.016	0.956
MLMI-Het	−0.822	0.062	0.063	0.947	0.360	0.019	0.020	0.958	−0.460	0.017	0.016	0.944
MLMI-SMC	−0.822	0.061	0.063	0.947	0.360	0.019	0.020	0.958	−0.460	0.017	0.016	0.946

NOTE.—20 percent missing data (MCAR) are introduced in each variable. We compare the true value of the parameters with the full data estimates and with the estimates handling missing data with CR, MI-noW, MI-W, MI-xW, MI-S, MLMI-Hom, MLMI-Het, and MLMI-SMC.

Table 4. MCS Analysis Results: Parameter Estimates and Associated Standard Error Estimates for the Four Fixed Effect Parameters of the Substantive Weighted Regression Model, β_0, β_1, β_2 and β_3

	β_0		β_1		β_2		β_3	
	Mean	SE	Mean	SE	Mean	SE	Mean	SE
CR	−0.683	0.051	0.292	0.013	0.101	0.033	−0.198	0.011
MI-noW	−0.749	0.047	0.313	0.012	0.090	0.030	−0.208	0.009
MI-W	−0.754	0.048	0.313	0.012	0.091	0.031	−0.208	0.009
MI-xW	−0.744	0.049	0.311	0.011	0.089	0.032	−0.208	0.009
MI-S	−0.703	0.083	0.309	0.021	0.074	0.049	−0.206	0.016
MLMI-Hom	−0.756	0.047	0.314	0.012	0.091	0.032	−0.208	0.009
MLMI-Het	−0.772	0.046	0.320	0.011	0.090	0.031	−0.206	0.009
MLMI-SMC	−0.756	0.049	0.314	0.012	0.092	0.031	−0.208	0.009

NOTE.—We compare CR estimates with the estimates handling missing data with CR, MI-noW, MI-W, MI-xW, MI-S, MLMI-Hom, MLMI-Het and MLMI-SMC.

Our substantive model is a weighted regression of the quantitative Bracken school readiness score on three explanatory variables: logarithm of family income, whether the child has hearing problems (1 = yes; 0 = no), and the number of siblings. The sampling weights are provided with the data; more detail on their derivation is given by Plewis (2007). We analyzed the complete records, and then we multiply imputed missing values using each of the methods presented in section 2. We used a burn in of 500 updates and imputed 20 datasets, updating the sampler 500 times between each imputation. As there are only nine weight strata, we use them to define the second level for the multilevel imputation method (i.e., all weights are the same within each stratum).

Table 4 summarizes the results. Compared with the CR analysis, only MI-S gives larger SEs, suggesting estimation in some strata is poor, so this approach may be less reliable here. The other best-performing methods from the simulation study (MLMI-Het, MLMI-SMC, and MI-xW) give similar results. Focusing on results from these three methods, compared with CR, they suggest (i) a $>$ 1 SE stronger positive effect of income on school readiness score (β_1); (ii) a slightly weaker effect of hearing problems (β_2), and (iii) a marginally stronger negative effect of a greater number of siblings (β_3).

6. DISCUSSION

In this article, we have reviewed some of the issues raised by using multiple imputation to impute missing values when the substantive analysis is a weighted model. This led us to propose a multilevel approach, where (i) the weights are used to form strata which define level-two, (ii) we include random intercepts, and (iii) we allow the variance structure of the variables to vary across strata. While random-effects models has been proposed previously as substantive models in order to shrink across weight strata (Elliott and Little 2000; Elliott 2007; Xia and Elliott 2016), to our knowledge, the use of a multilevel model, allowing associations between the variables to vary across weight strata in these settings, is new. Furthermore, while the possibility of using such models for the imputation of missing data has been suggested from a theoretical point of view (Zhou, Elliott, and Raghunathan 2016b), in this article, this strategy is evaluated through simulations and application of real data for the first time.

We have evaluated our approach in a series of simulation studies, finding encouraging results across all the scenarios. In applications, we may need to group weights in order to form strata. In this case, the approach is likely to perform best if the strata are relatively homogeneous.

Given our results, we believe that adopting our approach (either MLMI-Het or MLMI-SMC) addresses the issues raised by (Kim et al. 2006) and so renders their conclusion that “MI is not generally recommended for public use data files” unduly negative.

If our approach is adopted and the imputer and analyst are separate, we believe that those imputing data and subsequently releasing them for public use should also publish the imputation model so that users can see the structure that has been captured in the imputation model.

Across the scenarios we considered for MLMI-Het and MLMI-SMC, the empirical standard error was either close to or slightly larger than the model-based SE (obtained using Rubin’s rules), suggesting at worse the approach may be slightly conservative but still more efficient than CR (cf Meng, 1994).

Compared with the MI-xW approach, our multilevel approach does not rely on a linear association between the weights and the other variables; this can vary as dictated by the data across the weight strata. A further potential advantage of MLMI-SMC is that it can be combined with the approach outlined by Goldstein et al. (2014) to impute consistent with nonlinear relationships and interactions.

One possible issue with our proposed method is that with high dimensional data, incorporating all domains with their interaction might be complicated, even using our random effects to reduce the number of parameters. Because of this, an alternative multiple imputation method based on finite population Bayesian bootstrap has been recently proposed (Zhou, Elliott, and Raghunathan 2016a, 2016b); this method could have a potential advantage with large numbers of domains. We plan in the future to compare this method with our strategy to explore under which circumstances one is better than the other.

In this article, we focused on weights arising from simple random sampling. Extensions to consider more complex multistage sampling designs may be needed in applications. For example, our approach could have to be used to impute multilevel datasets. In this setting, the weight strata give rise to a cross-classified structure (Browne, Goldstein, and Rasbash 2001). In principle, this could be addressed simply by using a more complicated imputation model. This is a topic for future research.

For standard MI (i.e., when the substantive model is not weighted), our approach can also be used. In other words, we can estimate the probability of missing data and use these to form strata. If the weights are right, but the imputation model is wrong, preliminary simulations show this provides a degree of double-robustness.

The strategy we investigated in this article is tailored to handle missing items in surveys. We did not focus on the case of completely missing units, since we assume this will be addressed by the weights. This was also considered by Seaman et al. (2012), who recommended a two-stage combination of multiple imputation to handle item nonresponse and inverse probability weighting (IPW) to handle unit nonresponse.

To conclude, we have proposed and evaluated a multilevel multiple imputation approach for situations where the substantive analysis is weighted and found promising results. The approach described here is implemented in the R package jomo Quartagno and Carpenter (2014).

References

Bartlett, J. W., S. R. Seaman, I. R. White, and J. R. Carpenter (2015), "Multiple Imputation of Covariates by Fully Conditional Specification: Accommodating the Substantive Model," *Statistical Methods in Medical Research*, 24, 462–487.PMID: 24525487.

Beaumont, J., D. Haziza, and C. Bocci (2011), "On Variance Estimation under Auxiliary Value Imputation in Sample Surveys," *Statistica Sinica*, 21, 515–537.

Browne, W. J., H. Goldstein, and J. Rasbash (2001), "Multiple Membership Multiple Classification (Mmmc) Models," *Statistical Modelling*, 1, 103–124.

Carpenter, J., and M. Kenward (2013), *Multiple Imputation and Its Application*, Chichester: Wiley, ISBN: 978-0-470-74052-1.

Deville, J.-C., and C.-E. Särndal (1992), "Calibration Estimators in Survey Sampling," *Journal of the American Statistical Association*, 87, 376–382.

Elliott, M. R. (2007), "Bayesian Weight Trimming for Generalized Linear Regression Models," *Survey Methodology*, 33, 23–34.

Elliott, M. R., and R. J. A. Little (2000), "Model-Based Alternatives to Trimming Survey Weights," *Journal of Official Statistics*, 16, 191–209.

Fay, R. E. (1992), "When are Inferences from Multiple Imputation Valid?" Proceedings of the Survey Research Methods Section of the American Statistical Association, Washington, pp. 227–232.

———. (1993), "Valid Inferences from Imputed Survey Data," Proceedings of the Survey Research Methods Section of the American Statistical Association, Washington, pp. 41–48.

Goldstein, H., J. Carpenter, and W. Browne (2014), "Fitting Multilevel Multivariate Models with Missing Data in Responses and Covariates That May Include Interactions and Non-Linear Terms," *Journal of the Royal Statistical Society: Series A*, 177, 553–564.

Holt, D., and D. Elliot (1991), "Methods of Weighting for Unit Non-Response," *Journal of the Royal Statistical Society. Series D (The Statistician)*, 40, 333–342.

Kim, J. K., J. Michael Brick, W. A. Fuller, and G. Kalton (2006), "On the Bias of the Multiple-Imputation Variance Estimator in Survey Sampling," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68, 509–521.

Meng, X.-L. (1994), "Multiple-Imputation Inferences with Uncongenial Sources of Input: Rejoinder," *Statistical Science*, 9, 566–573.

Meng, X.-L., and M. Romero (2003), "Discussion: Efficiency and Self-Efficiency with Multiple Imputation Inference," *International Statistical Review*, 71, 607–618.

Morris, T. (2016), "Using Simulation Studies to Evaluate Statistical Methods in Stata: A Tutorial," London Stata Users Group Meeting 22nd, London, MRC Clinical Trials Unit at UCL, doi: 10.13140/RG.2.2.18161.6896.

Plewis, I. (2007), "Non-Response in a Birth Cohort Study: The Case of the Millennium Cohort Study," *International Journal of Social Research Methodology*, 10, 325–334.

Quartagno, M., and J. Carpenter (2014), *jomo: A package for Multilevel Joint Modelling Multiple Imputation*. R Package version 3.6.0.

Quartagno, M., and J. R. Carpenter (2016), "Multiple Imputation for Ipd Meta-Analysis: Allowing for Heterogeneity and Studies with Missing Covariates," *Statistics in Medicine*, 35, 2938–2954.

———. (2019), "Multiple Imputation for Discrete Data: Evaluation of the Joint Latent Normal Model," *Biometrical Journal* 61, 1003–1019. doi: 10.1002/bimj.201800222.

Quartagno, M., S. Grund, and J. R. Carpenter (2018), "jomo: A Flexible Package for Two-Level Level Joint Modelling Multiple Imputation," *R Journal*, in press).

Rao, J. N. K. (1996), "On Variance Estimation with Imputed Survey Data," *Journal of the American Statistical Association*, 91, 499–506.

Rao, J. N. K., and J. Shao (1992), "Jackknife Variance Estimation with Survey Data under Hot Deck Imputation," *Biometrika*, 79, 811–822.

Rubin, D. (1976), "Inference and Missing Data," *Biometrika*, 63, 581–592.

———. (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley.

Särndal, C., B. Swensson, and J. Wretman (1992), *Model Assisted Survey Sampling*, Springer Series in Statistics, New York: Springer.

Seaman, S. R., I. R. White, A. J. Copas, and L. Li (2012), "Combining Multiple Imputation and Inverse-Probability Weighting," *Biometrics*, 68, 129–137.

Xia, X., and M. R. Elliott (2016), "Weight Smoothing for Generalized Linear Models Using a Laplace Prior," *Journal of Official Statistics*, 32, 507–539.

Yucel, R. (2011), "Random-Covariances and Mixed-Effects Models for Imputing Multivariate Multilevel Continuous Data," *Statistical Modelling*, 11, 351–370.

Zhou, H., M. R. Elliott, and T. E. Raghunathan (2016a), "Synthetic Multiple-Imputation Procedure for Multistage Complex Samples," *Journal of Official Statistics*, 32, 231–256.

———. (2016b). "A Two-Step Semiparametric Method to Accommodate Sampling Weights in Multiple Imputation," *Biometrics*, 72, 242–252.