ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Applied Statistics
Series C

Appl. Statist. (2018)
67, *Part* 4, *pp.* 1071–1081

Bayesian models for weighted data with missing values: a bootstrap approach

Harvey Goldstein,

University of Bristol and London School of Hygiene and Tropical Medicine, UK

James Carpenter

London School of Hygiene and Tropical Medicine and University College London, UK

and Michael G. Kenward

London School of Hygiene and Tropical Medicine, UK

[Received May 2016. Final revision December 2017]

Summary. Many data sets, especially from surveys, are made available to users with weights. Where the derivation of such weights is known, this information can often be incorporated in the user's substantive model (model of interest). When the derivation is unknown, the established procedure is to carry out a weighted analysis. However, with non-trivial proportions of missing data this is inefficient and may be biased when data are not missing at random. Bayesian approaches provide a natural approach for the imputation of missing data, but it is unclear how to handle the weights. We propose a weighted bootstrap Markov chain Monte Carlo algorithm for estimation and inference. A simulation study shows that it has good inferential properties. We illustrate its utility with an analysis of data from the Millennium Cohort Study.

Keywords: Markov chain Monte Carlo sampling; Millennium Cohort Study; Missing data; Weighted bootstrap

1. Introduction

When analysing survey data, often the data are supplied with weights (e.g. representing the design, or to account for non-response). In addition, covariates in our substantive model may also have missing data so in such cases we require models that can simultaneously handle both weights and missing data. This has been an active research field, comprehensively summarized in chapters 7–9 of Molenberghs *et al.* (2015).

Goldstein *et al.* (2014) described an efficient, general, fully Bayesian procedure for handling missing data in a multilevel setting. This does not require multiply imputed data sets with the application of Rubin's rules. In this paper, we show how this approach can be extended to include weights, where we do not wish to condition, or cannot, our inferences on either the weights or the (possibly complex functions of) variables that are used in their estimation. This is important when we wish to estimate marginal quantities or population-averaged regression relationships which do not condition on either the weights or the variables that are used to

Address for correspondence: Harvey Goldstein, Graduate School of Education, University of Bristol, 35 Berkeley Square, Bristol, BS8 1JA, UK.
E-mail: h.goldstein@bristol.ac.uk

 0035–9254/18/671071

estimate the weights. We note that Bayesian approaches to the estimation of finite population parameters such as proportions utilize models that include functions of weights. Such models, however, do not address the issue of missing data nor the main aim of the present paper, which is concerned with inferences about parameters in the model of interest, which itself does not explicitly contain any function of the weights (see Chen *et al.* (2010) for a discussion).

Thus, unit weights in this paper derive from the study design where, for example, records are stratified and weights are used to recover population information, or where units are missing and we wish to assign weights on the basis of auxiliary data, such as interviewer observations or geographical information. Where individual items are missing we invoke the Bayesian imputation procedure that was given in Goldstein *et al.* (2014). The present paper derives an inferential procedure that simultaneously incorporates both individual unit weights and item missingness. We shall assume that values are missing either 'completely at random' or 'at random'. In the latter case we assume that covariates in the model of interest or auxiliary variables that enter the imputation part of the model are sufficient, when conditioned on (either explicitly or through the weights), so that the missingness mechanism can be adjusted for (Rubin, 1987).

For example, consider an analysis of data from wave 2 of the Millennium Birth Cohort Study (Plewis, 2007), when the children were approximately 3 years old. The study is a multidisciplinary research project following the lives of around 19000 children born in the UK in 2000–2001. The participants were selected from a random sample of electoral wards, that were stratified by a measure of deprivation and by a measure of the concentration of black and Asian families, so that areas of high deprivation and high concentrations of these ethnic minorities were over-represented. In our example, we seek to understand the relationship between the Bracken score of school readiness and the logarithm of family income at the wave 2 follow-up (14% missing values), whether the child has ever had hearing problems (14% missing values) and the number of siblings in the study child's household (no missing values). Overall there are 26% of records with at least one missing value. Further details are provided in Section 2 below.

Unfortunately, it is not obvious how directly to incorporate weights in a fully Bayesian framework, and a discussion of this issue can be found in Gelman (2007). The approach of the present paper is a hybrid one where a Bayesian model is used to handle the missing data and a bootstrap is used to incorporate the information from the weights. In this way, the posterior estimates incorporate the information in the weights without being conditioned on them.

Dong *et al.* (2014) proposed a procedure that has similarities to our own, although discussed in terms of generating 'synthetic' samples. They adopted a two-stage procedure. At the first stage they select a set of bootstrap samples from the observed data. Each of these has observations with associated weights. At the second stage a weighted (bootstrap) sample is chosen from each of these, resulting in an equally weighted sample that is representative of the population of interest. Appropriate models are then fitted to each of these samples and inference is carried out utilizing the first-stage set of bootstrap replicates. As they pointed out, missing values can be incorporated via standard multiple-imputation algorithms. Zhou *et al.* (2016) proposed essentially the same procedure, with the specific aim of handling missing data when weights are present. They used a chained equation method for multiple imputation and in their example their models of interest are fitted by maximum likelihood.

In the present paper, using a slightly different formulation but still involving a two-stage bootstrap, we extend these ideas by utilizing the fully Bayesian estimation method for missing data that was proposed by Goldstein *et al.* (2014). This simultaneously fits the substantive scientific model and imputes any missing data. It also allows full flexibility in fitting interaction and power terms in the substantive model of interest.

The outline for the paper is as follows. We begin by describing our motivating example in

Section 2. In Section 3 we describe our proposal in the absence of missing data. In Section 4 we build on Goldstein *et al.* (2014), utilizing the Bayesian approach to handle missing data (under the missingness at random assumption). We then explore the behaviour of our proposal by using simulations in Section 5 and apply it to our analysis of data from wave 2 of the Millennium Cohort Study. We conclude with a discussion.

2. The Millennium Cohort Study

The data consist of 13294 records from the Millennium Cohort Study at wave 2 (around 3 years of age) (see Plewis (2007) for sample details). We seek to understand factors affecting school readiness. Our substantive model is an additive linear regression of the square root of the Bracken school readiness score on three explanatory variables: $\log_e$(wave 2 family income), whether the child has hearing problems (1, yes; 0, no) and the number of siblings. The pattern of missing values is shown in Table 1. A description of the weights, derived from the sample design and taking account of attrition and non-response, can be found in Plewis (2007).

The estimates by using the 74% complete records, with and without weighting, are given in third and fourth columns respectively of Table 7 in Section 5.

3. A Hybrid Markov chain Monte Carlo algorithm for Bayesian estimation incorporating weights

We now outline the core of our proposed algorithm. Our illustrative substantive model of interest (which we later generalize) is, for a sample of size N,

$$y_i = x_i\beta + e_i, \qquad e_i \sim N(0, \sigma^2), \quad i = 1, \ldots, N, \tag{1}$$

where y_i is the response and x_i is a vector of covariates, and we write the response vector and covariate matrix as $Y = \{y_i\}$ and $X = \{x_i\}$.

As it stands model (1) can be fitted by a variety of algorithms. So that we can incorporate cases with missing values we fit this by using a Bayesian Markov chain Monte Carlo (MCMC) algorithm. This requires us to specify priors for the parameters and for present purposes we have chosen weak default priors. For the set of regression coefficients and the variance a Gibbs sampling step is used and in the next section when we introduce missing values a Metropolis step is used (see Goldstein *et al.* (2014)). As in our previous work, in the following models we have chosen the following independent default priors, namely for each of the regression coefficients diffuse priors $p(\beta) \propto 1$ and for the variance $\sigma^{-2} \sim \text{gamma}(0.001, 0.001)$.

In general, informative priors may be appropriate for the model of interest, but this raises no new issues and we shall not pursue this possibility. We note, however, that, in the context

Table 1. Missing value patterns in the wave 2 Millennium Cohort Study analysis: missing; observed

Missing values (%)	*Hearing*	*$\log_e$(income)*
74	Observed	Observed
12	Observed	Missing
12	Missing	Observed
2	Missing	Missing

of multilevel models where the number of higher level clusters is not large, then the choice of default priors for the corresponding variance parameters needs careful consideration (Gelman, 2006).

Now suppose that the data $\{y_i,\ x_i\},\ i=1,\ldots,N$, are supplied with corresponding weights $W=\{w_i\}$ which are inversely proportional to the sampling probabilities Thus, standard weighted regression will give valid estimates of population parameters estimates, though not valid estimates of other inferential quantities such as standard errors. Hence, in frequentist inference, sandwich variance estimators are often used.

3.1. Using the bootstrap for inference: a two-stage procedure

To introduce our proposal we first consider an intuitive one-stage bootstrap approach that is outlined in steps A1 and A2 below and then discuss its limitations and why a two-stage bootstrap is required.

Consider a weighted non-parametric bootstrap that draws bootstrap samples where records are selected by using the weights. This then provides a 'representative' sample which can fit into a Bayesian framework where missing data are handled, with inferences made from the set of boostrap sample estimates. The following steps are used, where step A2 uses the MCMC algorithm of Goldstein *et al.* (2014), as follows.

Step A1: select a sample with replacement from $\{y_i, x_i\},\ i=1,\ldots,N$, by using the weights W.
Step A2: perform one iteration of the MCMC chain for the parameters.

Steps A1 and A2 are repeated for r iterations. If there are no missing values, then the resulting chain gives valid inference for the regression parameter means. This follows because the weights have been chosen so that, for each weighted sample, the expectation of the regression parameters is equal to the population values.

However, if there are missing data, then at each step A2 we need additionally to generate Bayesian draws for the missing values. At step A2 we therefore need a suitable burn-in s for these. We consider the choice of r, s and b in Section 3.2 below.

Although we might suppose that steps A1 and A2 alone would be sufficient for valid inference, in fact this is not so; the weighting causes underestimation of standard errors (which we have illustrated empirically; see the second column of Table 4 in Section 4). Our intuition for this is as follows: when the sampling probabilities (and hence the weights) are a function of the dependent or response variable, the elements of a weighted bootstrap sample are independent, conditional on the weights (e.g. within strata defined by weights). Our model of interest, however, does not condition on the weights and in effect marginalizes over the weights, so that the elements will no longer be independent. To address this inferential problem, we propose to embed steps A1 and A2 in an outer level, equally weighted, bootstrap that thus ensures independence, as follows.

Step B1: select an equally weighted sample of size N with replacement from $\{y_i,\ x_i\},\ i=1,\ldots,N$ (keeping the weight for unit i).
Step B2: apply steps A1 and A2.

We then repeat steps B1 and B2 m times. We shall refer to the two stages, 'B' and 'A', as respectively first- and second-stage bootstraps.

In conventional MCMC sampling, we typically run steps A1 and A2 a large number, K, of times to obtain parameter estimates. We propose instead to save computational time by choosing r and m so that $mr = K$. Our parameter estimates and bootstrap standard errors are calculated as follows.

For each cycle $i = 1, \ldots, m$, of stage B, we obtain $j = 1, \ldots, r$ draws of the vector of model parameters, denoted $\{\theta_{ij}\}$.

Our point estimate of θ is then

$$\hat{\theta} = \frac{1}{mr} \sum_{ij} \theta_{ij},$$

with estimated covariance matrix

$$\hat{\Sigma} = \frac{1}{m} \sum_{i} \tilde{\theta}_{i.}^{\mathrm{T}} \tilde{\theta}_{i.},$$

where $\hat{\theta}_{i.} = (1/r)\Sigma_j \theta_{ij}$, $\tilde{\theta}_i = \hat{\theta}_{i.} - \hat{\theta}$ and for inference we either assume that $\hat{\theta} \sim N(\theta, \hat{\Sigma})$ or choose m sufficiently large that we can use bootstrap estimates of the quantiles to derive interval estimates.

3.2. Choosing m, r and s

For the bootstrap estimate of standard errors, values of m in the range 50–200 have been suggested (Efron and Tibshirani (1993), page 14), although more will be needed for estimating a covariance matrix. We therefore evaluated $m = 100$ in our simulation study below.

In the inner bootstrap, we initially chose $r = 25$ to give an overall chain length of $K = mr = 2500$ for estimating the parameter means. In the simulation study we have explored whether inference was compromised by smaller values. Lastly, with missing data, we need to choose s, the burn-in at each step A2, before drawing the values to propagate through the algorithm. Since, after the initial burn-in, the parameter estimates are stable, it seems intuitively that s can be relatively small; we compare $s = 1$ and $s = 5$ below.

3.3. Further comments

We note that this procedure can be used for any model for which an MCMC algorithm exists; this includes multilevel and cross-classified data. Similarly, any pattern of missing data can be accommodated. To handle non-linear relationships and interactions correctly when data are missing we use the fully Bayesian missing data model that was suggested by Goldstein *et al.* (2014), as described above.

To preserve the properties of the bootstrap, and to ensure consistent estimates, where the same record with a missing value is sampled more than once in a stage 2 draw, the imputed values should be constrained to be equal. This is most conveniently done by imputing for the first occurrence of the record and copying the imputed value to the second and subsequent occurrences, and is implemented in our software.

Starting values for the missing values need to be given at each step A2, and good choices will considerably speed up the algorithm. Our proposal, that we use in our examples, is that the algorithm is initiated by estimating the predictive distribution using complete cases in the weighted selection drawn the first time that step A2 is executed.

4. Simulation study

We evaluate our proposals for m, r and s by using a simulation study where the weights are related to population strata means. We show a sequence of analyses where we vary the number of bootstraps, the sample size and the burn-in period for the missing values.

We consider 10 strata. In each stratum ($j = 1, \ldots, 10$) define a multivariate normal distribution

$$\begin{pmatrix} y_j \\ X_j \end{pmatrix} \sim N(\mu_j, \Omega_j)$$

where

$$\mu_j = \left(\frac{j}{10}, \frac{j}{10}, \frac{j}{10}, \frac{j}{10}, \frac{j}{10}\right).$$

We set up the following population model of interest with four explanatory variables $y = X\beta + e$:

$$y = \{y_j\}, \qquad X = \{X_j\}, \qquad \Omega_j = \frac{1}{j}\begin{pmatrix} 2 & & & & \\ 1 & 2.5 & & & \\ 1 & 1.5 & 2.5 & & \\ 1 & 1.5 & 1.5 & 2.5 & \\ 1 & 1.5 & 1.5 & 1.5 & 2.5 \end{pmatrix} \tag{2}$$

so that the relationship varies across strata.

Given the definition of the population, we can consider drawing a simple random sample of any given size. We shall draw a very large such sample ($n = 10^6$) to obtain good estimates of our population parameters. Thus, given the parameter values in model (6) for stratum $j = 1$, the mean population coefficient is estimated as 0.14 and for stratum 10 it is 0.23. For the overall model the population coefficient estimates are $\beta = (0.175, 0.174, 0.173, 0.174)$. The basic sample size for analysis is 1000 and we draw equal numbers from each stratum so that the weights are proportional to $1/\sqrt{j}$. We also study the effect of increasing this sample size.

We note that the values of the population units have a distribution that is a weighted normal mixture. Thus the marginal conditional relationship in our model is strictly misspecified. For our purposes, however, we are interested in this model as a summary and our purpose is to explore whether the algorithm proposed, using the same distributional assumptions, yields the same

Table 2. Simulation results based on 100 simulations for sample size $N = 1000$ by using bootstrap weighting with MCMC estimates with stratum sizes proportional to $1/\sqrt{j}$†

Parameter	*Values for s = 1*	*Values for s = 5*
β_1	16.3 (93.0)	4.3 (94.0)
β_2	2.7 (92.0)	0.4 (92.0)
β_3	1.9 (97.0)	0.7 (91.0)
β_4	3.1 (97.0)	3.0 (95.0)
$\bar{\beta}$	3.5 (94.8)	1.9 (93.0)
σ_e^2	1.2 (97.0)	2.9 (93.0)

†For each sample the (absolute) percentage relative bias and 95% interval percentage coverage estimate (in parentheses) are shown. (The coverage estimates are based on a normal assumption for the parameter estimates. For the variance, with a mean of approximately 0.6 and $m = 100$ this will provide a reasonable approximation.) The number of second-stage bootstrap-weighted selections for each (of m) first-stage equally weighted bootstrap is 25 with 100 first-stage bootstraps. An overall burn-in of 500 is used. $\bar{\beta}$ is the mean coverage and bias taking account of the sign. 15% of the values in variables 2, 3 and 4 are independently set to be randomly missing. Second-stage missing data burn-in (s) values are shown. Starting values for missing values are chosen as the means over the non-missing values and a normal proposal distribution uses the conditional variance.

Table 3. Estimates of percentage relative bias, based on simulation results for sample size 1000 and various values of r with $s = 5$ and other values as in Table 2 (95% interval percentage coverage estimates)

Parameter	*Values for r = 5*	*Values for r = 10*
β_1	5.4 (97.0)	1.6 (97.0)
β_2	5.5 (96.0)	0.7 (94.0)
β_3	6.0 (100.0)	9.6 (91.0)
β_4	0.2 (96.0)	0.7 (96.0)
$\bar{\boldsymbol{\beta}}$	1.6 (97.3)	2.0 (95.0)
σ_e^2	2.6 (97.0)	2.3 (90.0)

Table 4. Simulation results for various sample sizes N with $s = 5$ and $r = 10$ and other values as in Table 2: percentage relative bias (95% interval percentage coverage estimates)†

Parameter	*Results for N = 1000, no first-stage bootstrap (m = 1)*	*Results for N = 1000, m = 100*	*Results for N = 2000, m = 100*
β_1	1.7 (84.0)	0.2 (95.0)	0.0 (96.0)
β_2	1.4 (89.0)	4.9 (92.0)	5.4 (98.0)
β_3	1.9 (88.0)	1.2 (92.0)	2.1 (94.0)
β_4	7.0 (90.0)	2.3 (96.0)	0.0 (94.0)
$\bar{\boldsymbol{\beta}}$	1.2 (87.8)	2.1 (93.8)	1.9 (95.5)
σ_e^2	2.7 (84.0)	1.3 (95.0)	1.7 (94.0)

†Also shown in the second column are results from omitting the first-stage bootstrap.

summary estimates as would be obtained by fitting the population model to a simple random sample from the population.

Table 2 shows results for two different missing value burn-in sizes, with 100 first-stage bootstraps, a sample size of 1000 and 25 second-stage weighted bootstraps for each first-stage bootstrap. We randomly and independently choose about 15% of values for variables 2–4 to be missing and on average about 40% of records have at least one missing value.

Table 2 suggests that a burn-in of five for the missing covariate values has a small bias of about 2% for the fixed effects and 3% for the variance, with reasonable coverage and even a burn-in of one shows acceptable bias and coverage. Further calculations with greater values for s suggest little further improvement. A suitable value for this burn-in will depend on the data in any application and in practice different values may be tried and this is an area for further research.

In Table 3 we look at varying the number of second-stage replications r.

We see that we appear to obtain acceptable bias and coverage even for $r = 5$ and in subsequent analyses we shall in fact use a conservative value of 10.

In Table 4 we look at varying the sample size.

In the third and fourth columns, for both sample sizes the average bias for the fixed parameters is about 2%, with a smaller bias for the variance. The coverage is good overall. The second column shows the relatively poor coverage that occurs when the first-stage bootstrap is omitted.

Table 5. Simulation results for two sample sizes N, using bootstrap weighting with MCMC estimates with stratum sizes proportional to $1/\sqrt{j}$: binary response model with probit link†

Parameter	*Values for N = 1000*	*Values for N = 2000*
β_1	0.9 (95.0)	1.6 (100)
β_2	11.1 (93.0)	5.5 (96.0)
β_3	13.1 (97.0)	0.4 (96.0)
β_4	3.0 (97.0)	5.3 (96.0)
$\bar{\boldsymbol{\beta}}$	6.6 (95.5)	3.0 (97.0)

†The absolute percentage relative biases (95% interval percentage coverage) are shown. The total number of second-stage bootstrap-weighted selections for each first-stage equally weighted bootstrap is $r = 10$, with $m = 100$ first-stage bootstraps. The overall burn-in is 500 and the burn-in for missing values is $s = 5$.

Table 6. Absolute percentage mean bias (over the four regression coefficients) and coverage (in parentheses) followed by bias (coverage) for the residual variance, by sample size and number of strata†

Number of strata	*Results for the following sample sizes:*		
	1000	*3000*	*5000*
10	0.5 (94.8), 3.0 (93.0)	0.1 (94.0), 1.4 (94.0)	0.4 (92.3), 0.9 (92.0)
50	0.1 (89.8), 6.8 (83.0)	0.1 (95.3), 2.1 (93.0)	0.0 (93.8), 0.8 (97.0)
100	0.0 (94.5), 12.2 (78.0)	0.0 (93.0), 5.1 (91.0)	0.0 (93.0), 1.2 (95.0)

†100 simulations, no missing data; $r = 10$; $s = 1$ since there are no missing data.

We now dichotomize our response so that the simulated responses become 1 if greater than 0 and 0 otherwise. As in the case of missing predictor variable values, if the same stage 2 record is sampled more than once, the sampled, latent normal, responses are constrained to be equal.

Table 5 displays the results for 100 simulations by using the same specification as before.

We see that the performance of the algorithm is somewhat worse for a sample size of 1000 compared with the case of a normal response, although the coverage remains good. For a sample size of 2000 the results are comparable with those found with a normal response. In the case of binary responses, for a given sample size, we have less information than for normal responses so a poorer performance in terms of bias and coverage is to be expected. Further small-scale simulations show decreasing amounts of bias as the sample size increases further.

The number of distinct weights, in our example the number of strata, as well as their heterogeneity can affect the bias and coverage, irrespectively of whether there are any missing data. We illustrate this in Table 6 where we carry out simulations as above, with no missing data, but varying the sample sizes and the number of strata.

We see that, for the sample size of 1000, distributed across 50 and 100 strata there is a bias, which is typically negative, for the variance and poor coverage, although the fixed coefficients

Table 7. Different estimation procedures for the Millennium Cohort Study data†

Parameter	*Results for MCMC sampling ($r = 10$; $m = 100$; missing value burn-in, 5)*	*Results for weighted ordinary least squares with sandwich estimates ($N = 9866$)*	*Results for equally weighted ordinary least squares with sandwich estimates ($N = 9866$)*
β_0	−0.740 (0.052)	−0.729 (0.048)	−0.749 (0.045)
β_1	0.314 (0.012)	0.307 (0.012)	0.304 (0.011)
β_2	0.077 (0.031)	0.088 (0.030)	0.088 (0.028)
β_3	−0.209 (0.009)	−0.209 (0.009)	−0.190 (0.010)
σ_e^2	0.797 (0.012)	0.809 (0.014)	0.815 (0.014)

†MCMC chain burn-in, 500; number of iterations, 2500 with diffuse priors. Standard errors are in parentheses; $N = 13294$.

generally exhibit acceptable bias and adequate coverage. For the variance this improves with the larger sample sizes. Increasing r has little effect on the bias or coverage.

5. Analysis of Millennium Cohort Study data

The data consist of 13294 records from the Millenium Cohort Study at wave 2 (around 3 years of age) (see Plewis (2007) for sample details) and we fit an additive normal linear regression model with an intercept β_0 and the three further explanatory variables and missing data pattern as described in Section 2. A description of the weights, derived from the sample design and taking account of attrition and non-response, can be found in Plewis (2007). In Table 7 we present results where the second column fits the model by using the two-stage weighted bootstrap and imputing missing values, the third column gives estimates from a standard weighted analysis for complete cases with sandwich estimators and the fourth is for a model where the weights are assumed to be equal.

For the complete-case analysis an equally weighted fit gives some slightly different estimates compared with the weighted analysis for complete cases only. For the bootstrap procedure the estimates differ little from the complete-cases estimates with similar standard errors. Although the coefficient for whether the child has ever had hearing problems (β_2) is somewhat smaller, the associated standard error estimate is large. The bootstrap procedure took approximately 20 h by using non-optimized test code on a 2.4 GHz personal computer with an X-64-based processor under Windows 7.

6. Discussion

Incorporating Bayesian estimation via MCMC sampling within a bootstrap allows us to incorporate efficient methods for missing data, especially in the case where there are interaction terms in the model of interest (Goldstein *et al.*, 2014), and the resulting inferences utilize general properties of the bootstrap and (as our simulations show) have good frequentist properties. Our simulation results provide indications of likely biases and coverage estimates. Where the range of weights is not too large or where the sample size is not small the suggested procedure produces good coverage and small bias.

Where the range of weights (in our case the number of strata) is large we would not expect our procedure to work well, as we show in Table 6. In such cases the procedures suggested by Carpenter and Kenward (2013), chapter 11, based on weighted multiple imputation can be used

and further work extending these is currently being pursued. Alternatively, as an approximation, the weights can be divided into a small number of categories with the mean weight in each category used. We note, however, that, in principle, our procedure will automatically deal with the case where every record in a data set has a unique weight.

Although in our example the weights are a composite of stratum design weights and initial non-response weights, our simulation deals only with the former. We would argue that any weighting for non-response that is derived by using additional information, e.g. through propensity matching, is generally better dealt with by incorporating such information via auxiliary variables within the imputation model. An area for further exploration, however, is the extent to which this is so.

In the case of multilevel (hierarchical) data, it seems (Carpenter *et al.*, 2003) that a general non-parametric bootstrap does not exist. Thus an alternative is either a semiparametric or fully parametric bootstrap. It is not clear how one would carry out a fully parametric bootstrap in our proposed framework, since there appears to be no way to associate the weights with the sampled random effects. For a semiparametric bootstrap, in the case of a two-level model, we can replace steps B1 and B2 as follows to obtain a modified bootstrap sample.

Step B1: we carry out an unweighted MCMC model fit, incorporating missing data. Residuals at level 1 and level 2 are estimated.
Step B2: using the reflation procedure (Carpenter *et al.* (2003) and Goldstein (2011), chapter 3) we obtain a modified set of random effects, corresponding modified responses and thus a bootstrap data set for passing to the next stage of the algorithm.

Where the model response is non-normal the algorithm essentially consists of the same steps but with some small modifications. Specifically, in the case of binary response models, we utilize for the model of interest a probit link function with a latent normal specification as outlined for the case where there are missing values for binary predictors. This therefore entails an extra step in the algorithm that samples a standard normal variable given the binary response and current values of the parameters.

When burning in for the missing values of the explanatory variables it is generally more efficient to use the binomial likelihood that is associated with the model of interest rather than sampling a new set of latent normal variables at each burn-in iteration. Further work around this issue is planned, especially investigating multilevel structures. Software for carrying out the computations has been written in MATLAB (Mathworks, 2015), which is available from the first author. The estimation is computer intensive and ways of increasing the computational efficiency are being studied with a view to incorporating an efficient implementation in the package STATJR (STATJR, 2015).

Finally, we note that for the missing data we could use a multiple-imputation approach rather than the fully Bayesian approach that was adopted here, although this will not deal properly with interaction terms or multilevel data with missingness at higher levels. Nevertheless, where these considerations do not apply, it may be useful, e.g. for purposes of secondary analyses using standard software packages. It may also be faster in that the step for sampling the parameters of interest is absent from the chain iterations.

Acknowledgements

We are very grateful to the referees for their helpful comments.

James Carpenter is grateful for support from the Medical Research Council, grant MC_UU_12023/21.

References

Carpenter, J. R., Goldstein, H. and Rasbash, J. (2003) A novel bootstrap procedure for assessing the relationship between class size and achievement. *Appl. Statist.*, **52**, 431–443.

Carpenter J. R. and Kenward M. G. (2013) *Multiple Imputation and Its Application*. Chichester: Wiley.

Chen, Q., Elliott, M. R. and Little, R. J. A. (2010) Bayesian penalised spline model-based inference for finite population proportion in unequal probability sampling. *Surv. Methodol.*, **36**, 23–34.

Dong, Q., Elliot, M. R. and Raghunathan, T. E. (2014) A nonparametric method to generate synthetic populations to adjust for complex sampling design features. *Surv. Methodol.*, **40**, 29–46.

Efron, B. and Tibshirani, R. J. (1993) *An Introduction to the Bootstrap*. New York: Chapman and Hall.

Gelman, A. (2006) Prior distributions for variance parameters in hierarchical models. *Baysn Anal.*, **1**, 515–533.

Gelman, A. (2007) Struggles with survey weighting and regression modeling. *Statist. Sci.*, **22**, 153–164.

Goldstein, H. (2011) *Multilevel Statistical Models*, 4th edn. Chichester: Wiley.

Goldstein, H., Carpenter, J. R. and Browne, W. J. (2014) Fitting multilevel multivariate models with missing data in responses and covariates that may include interactions and non-linear terms. *J. R. Statist. Soc.* A, **177**, 553–564.

Mathworks (2015) *MATLAB*. Natick: Mathworks. (Available from `http://uk.mathworks.com/products/matlab/`.)

Molenberghs, G., Fitzmaurice, G., Kenward, M. G., Tsiatis, A. and Verbeke, G. (2015) *Handbook of Missing Data Methodology*. New York: CRC Press.

Plewis, I. (2007) Non-response in a birth cohort study: the case of the millennium cohort study. *Int. J. Socl Res. Methodol.*, **10**, 325–334.

Rubin, D. B. (1987) *Multiple Imputation for Non Response in Surveys*. Chichester: Wiley.

STATJR (2015) STATJR. Centre for Multilevel Modelling, University of Bristol, Bristol. (Available from `http://www.bristol.ac.uk/cmm/software/statjr/`.)

Zhou, H., Elliot, M. R. and Raghunathan, T. E. (2016) A two-step semiparametric method to accommodate sampling weights in multiple imputation. *Biometrics*, **72**, 242–252.