Biometrika (2017), **104**, 1, *pp.* 221–228
Printed in Great Britain

doi: 10.1093/biomet/asw058
Advance Access publication 3 January 2017

Miscellanea

A note on multiple imputation under complex sampling

By J. K. KIM

Department of Statistics, Iowa State University, 1208 Snedecor Hall, Ames, Iowa 50011, U.S.A.
jkim@iastate.edu

and S. YANG

Department of Statistics, North Carolina State University, 2311 Stinson Drive, Campus Box 8203, Raleigh, North Carolina 27606, U.S.A.
syang24@ncsu.edu

Summary

Multiple imputation is popular for handling item nonresponse in survey sampling. Current multiple imputation techniques with complex survey data assume that the sampling design is ignorable. In this paper, we propose a new multiple imputation procedure for parametric inference without this assumption. Instead of using the sample-data likelihood, we use the sampling distribution of the pseudo maximum likelihood estimator to derive the posterior distribution of the parameters. The asymptotic properties of the proposed method are investigated. A simulation study confirms that the new procedure provides unbiased point estimation and valid confidence intervals with correct coverage properties whether or not the sampling design is ignorable.

Some key words: Approximate Bayesian computation; Bayesian inference; Informative sampling; Item nonresponse; Pseudo maximum likelihood estimator.

1. Introduction

Item nonresponse is frequently encountered in survey sampling and imputation is a popular tool for handling item nonresponse. Multiple imputation, proposed by Rubin (1987, 1996) and further extended by Rubin & Schenker (1986), has been used as a general method for estimating the precision of sample estimates in the presence of imputed values. The technique has also been applied to a number of large-scale surveys (Schenker et al., 2006). See Little & Rubin (2002) for a comprehensive overview.

An attractive feature of multiple imputation is that complete-data analyses can be applied straightforwardly to the imputed datasets and these multiple results are summarized by an easy-to-implement combining rule for inference. In multiple imputation, M completed datasets are created. For each dataset, the estimate $\hat{\theta}_{I(k)}$ of a population parameter θ $(k = 1, \ldots, M)$ is computed. The overall estimate is the average of these estimates, $\hat{\theta}_{\text{MI}} = M^{-1} \sum_{k=1}^{M} \hat{\theta}_{I(k)}$. The multiple imputation variance estimator of $\hat{\theta}_{\text{MI}}$ is $\hat{V}_{\text{MI}} = U_M + \left(1 + M^{-1}\right) B_M$, where $U_M = M^{-1} \sum_{k=1}^{M} \hat{V}_{I(k)}$ accounts for the within-imputation variance, $B_M = (M-1)^{-1} \sum_{k=1}^{M} (\hat{\theta}_{I(k)} - \hat{\theta}_{\text{MI}})^2$ accounts for the between-imputation variance, and $\hat{V}_{I(k)}$ is the variance estimator computed from the kth dataset, treating imputed values as if they were observed values.

Current practice for multiple imputation assumes that the sampling design is ignorable (Rubin, 1976) or noninformative (Pfeffermann & Sverchkov, 1999), so the sample distribution equals the population distribution. The asymptotic properties of multiple imputation under ignorable sampling designs have been investigated extensively, for example by Rubin (1987), Schenker & Welsh (1988), Kott (1995), Wang & Robins (1998), Kim et al. (2006), and Yang & Kim (2016). When the sampling design is nonignorable, however, the sample distribution may not equal the population distribution and ignoring the sampling design can lead to biased estimation (Scott, 1977; Pfeffermann, 1993; Pfeffermann et al., 1998). In particular, the multiple imputation estimator can be biased even when a design-unbiased analysis method is applied to the imputed data. The usual recommendation in this case is to augment the imputation model by including the design information in the model, so that the sampling design becomes ignorable under the augmented model. While such an augmented model approach is promising, all the information related to the sampling design is not always available to data analysts. Also, the sample obtained from complex sampling often undergoes calibration, nonresponse adjustment, poststratification, raking and weight trimming adjustment. It is not clear how to incorporate such components into the imputation model.

In this paper, we develop a new multiple imputation method for complex sampling that does not use the augmented model approach. To achieve this, we propose a new data augmentation algorithm to carry out Bayesian inference, where we use the sampling distribution of the pseudo maximum likelihood estimator to derive the posterior distribution of the parameters. The proposed approach is a version of approximate Bayesian computation using the posterior distribution of parameters conditional on summary statistics (Fearnhead & Prangle, 2012; Soubeyrand & Haon-Lasportes, 2015). Our approach differs from the traditional Bayesian imputation approach in that we do not necessarily specify the full sample-data likelihood. The proposed multiple imputation method is similar in spirit to the calibrated Bayesian approach (Little, 2012) as the resulting inference is based on a design-based inference framework while maintaining the advantage of Bayesian analysis.

2. Methodology

Suppose that the finite population $\mathcal{F}_N = \{(x_i, y_i) : i \in U_N\}$ with $U_N = \{1, \ldots, N\}$ is a random sample from an infinite population ζ with joint density $f(y \mid x)f(x)$, where $f(y \mid x) = f(y \mid x; \theta)$ for some $\theta \in \Theta \subset \mathbb{R}^d$ and $f(x)$ is completely unspecified. From the realized finite population, we select a sample $A \subset U_N$ by a probability sampling design. Each unit in the sample is associated with a sampling weight ω_i, which is determined by some design variables. Often, only the sampling weight is made available to the public for confidentiality reasons. The sampling weight can be obtained from complex weighting such as poststratification. We assume that $\hat{Y}_n = \sum_{i \in A} w_i y_i$ is asymptotically design-unbiased for $Y = \sum_{i=1}^{N} y_i$ and that the variance estimator, $\hat{V}_n = \sum_{i \in A} \sum_{j \in A} \Omega_{ij} y_i y_j$, is design-consistent for $\text{var}(\hat{Y}_n)$.

From the sample, suppose that x_i is always observed but y_i is subject to missingness. Let δ_i take the value 1 if y_i is observed and 0 otherwise. Let $X_n = \{x_i : i \in A\}$, $Y_n = \{y_i : i \in A\}$, and $Y_n = (Y_{\text{obs}}, Y_{\text{mis}})$, where Y_{obs} and Y_{mis} are the observed and missing parts of Y_n, respectively. We assume missingness at random in the sense that $\text{pr}(y_i \in B \mid x_i, \delta_i = 1) = \text{pr}(y_i \in B \mid x_i)$ for any measurable set B and for all x_i, which is assumed at the population level and can be called population missingness at random according to Berg et al. (2016). The missingness mechanism here is conceptualized as a function of inherent characteristics of the units in the population and it does not depend on the sample design. If the missingness mechanism is viewed as a process amenable to scientific examination (Schafer, 1997), it is natural to define such mechanisms at the population rather than the sample level. This would hold if x_i contains all predictors for both y_i and δ_i. In practice, we may rely on subject-matter knowledge to collect a rich set of variables so that this assumption holds at least approximately. Extension of our method to population missingness not at random will be a topic of future study. Under population missingness at random, the imputed estimator $\hat{Y}_I = \sum_{i \in A} w_i \{\delta_i y_i + (1 - \delta_i) y_i^*\}$ is approximately design-model unbiased for Y if y_i^*, the imputed value for y_i, satisfies $E(y_i^* - y_i \mid x_i) = 0$.

In classical multiple imputation, the imputed values are generated as follows:

Step 1. Generate θ^* from the posterior density,

$$\theta^* \sim p(\theta \mid X_n, Y_{\text{obs}}) = \frac{\int L_s(\theta \mid X_n, Y_n)\pi(\theta)\, \mathrm{d}Y_{\text{mis}}}{\int \int L_s(\theta \mid X_n, Y_n)\pi(\theta)\, \mathrm{d}Y_{\text{mis}}\, \mathrm{d}\theta}, \tag{1}$$

where $L_s(\theta \mid X_n, Y_n)$ is the sample-data likelihood of (X_n, Y_n) viewed as the function of θ and $\pi(\theta)$ is the prior density of θ.

Step 2. For each unit with $\delta_i = 0$, generate y_i from the imputation model evaluated at θ^*, $y_i^* \sim f(y_i \mid x_i; \theta^*)$.

To generate θ^* from (1), the data augmentation algorithm (Tanner & Wong, 1987) can be used, which iterates the following two steps until convergence:

I-step. Given the parameter value θ^*, generate $y_i^* \sim f(y_i \mid x_i; \theta^*)$ for $\delta_i = 0$.

P-step. Given the imputed values y_i^*, generate

$$\theta^* \sim p(\theta \mid X_n, Y_n^*) = \frac{L_s(\theta \mid X_n, Y_n^*)\pi(\theta)}{\int L_s(\theta \mid X_n, Y_n^*)\pi(\theta)\,\mathrm{d}\theta}, \tag{2}$$

where $Y_n^* = (Y_{\text{obs}}, Y_{\text{mis}}^*)$ uses the imputed values generated from the I-step.

Here the I-step is the imputation step and the P-step is the posterior sampling step. Under an ignorable sampling design, the sample-data likelihood can be based on the population model, i.e., $L_s(\theta \mid X_n, Y_n) = \prod_{i\in A} f(y_i \mid x_i; \theta)$. To achieve sampling design ignorability, all design features should be built into the imputation model. Researchers have investigated multiple imputation for different sampling schemes, and argued that the imputation model should include random cluster effects for cluster samples, fixed stratum effects for stratified samples, and the size variable for probability-proportional-to-size sampling (Rubin, 1996; Yuan & Little, 2007; Chen et al., 2010). While such an augmented model approach can make the sampling design ignorable in principle, specifying the correct augmented model is difficult (Reiter et al., 2006). Furthermore, multiple imputation using the augmented model approach can still result in biased estimation because the missingness mechanism can become nonignorable if the augmented covariates share an unobserved common cause with the missingness mechanism (Berg et al., 2016). See Setting II in the simulation study.

We consider an alternative approach that does not use the sample-data likelihood $L_s(\theta)$ in (1) and does not require the sampling mechanism to be ignorable. To describe the proposed approach, we first consider the pseudo maximum likelihood estimator of θ in $f(y \mid x; \theta)$ under complete response by solving

$$\sum_{i\in A} w_i S(\theta; x_i, y_i) = 0, \tag{3}$$

where $S(\theta; x, y) = \partial \log f(y \mid x; \theta)/\partial\theta$ is the score function of θ. The pseudo maximum likelihood estimator is widely adopted in survey sampling, as it can provide consistent parameter estimators even under nonignorable sampling designs (Godambe & Thompson, 1986; Pfeffermann, 1993; Chambers & Skinner, 2003, Ch. 2; Korn & Graubard, 2011, Ch. 3). The sampling variance of $\hat\theta$ is estimated by the linearization formula (Binder, 1983; Kim & Park, 2010). Under the regularity conditions discussed in Fuller (2009), $\mathrm{var}(\hat\theta \mid \theta)^{-1/2}(\hat\theta - \theta) \to N(0, I)$ in distribution as $n \to \infty$, where $\mathrm{var}(\hat\theta \mid \theta)$ is the covariance matrix of $\hat\theta$.

Next, using the idea of approximate Bayesian computation, we can construct a new posterior density of θ conditional on $\hat\theta = \hat\theta(X_n, Y_n)$, which can be written as

$$p_g(\theta \mid X_n, Y_n) = \frac{g(\hat\theta \mid \theta)\pi(\theta)}{\int g(\hat\theta \mid \theta)\pi(\theta)\,\mathrm{d}\theta}, \tag{4}$$

where g is the sampling distribution of $\hat\theta$ and $\pi(\theta)$ is a prior density of θ. Under the existence of missing data, instead of using (2), we propose using the following P-step in the data augmentation algorithm:

New P-step. Given the imputed values y_i^*, generate

$$\theta^* \sim p_g(\theta \mid X_n, Y_n^*) = \frac{g(\hat{\theta}^* \mid \theta)\pi(\theta)}{\int g(\hat{\theta}^* \mid \theta)\pi(\theta)\,\mathrm{d}\theta}, \tag{5}$$

where $\hat{\theta}^* = \hat{\theta}(X_n, Y_n^*)$ is the solution to (3) using the imputed values generated from the I-step.

Roughly speaking, in the original P-step, the likelihood function of θ is based on the sample data (X_n, Y_n), whereas in the new P-step, the likelihood function of θ is replaced by the sampling distribution of the pseudo maximum likelihood estimator $\hat{\theta}(X_n, Y_n)$. The new data augmentation algorithm implies that the posterior density based on the observed data is

$$p_g(\theta | X_n, Y_{\text{obs}}) = \frac{\int g(\hat{\theta}|\theta)\pi(\theta)\,\mathrm{d}Y_{\text{mis}}}{\int\int g(\hat{\theta}|\theta)\pi(\theta)\,\mathrm{d}Y_{\text{mis}}\,\mathrm{d}\theta}. \tag{6}$$

Once M imputed datasets are generated from the above data augmentation method, we can apply Rubin's formula to combine estimates from each dataset.

Remark 1. Under an ignorable sampling design, the posterior density $p_g(\theta \mid X_n, Y_n)$ in (4) equals the classical posterior density in (2). To see this, let $l_s(\theta) = n^{-1}\sum_{i \in A} \log f(y_i \mid x_i; \theta)$ be the loglikelihood of θ. Since the sampling design is ignorable, the maximum likelihood estimator $\hat{\theta}$ obtained from $l_s(\theta)$ is consistent for θ. By a Taylor expansion,

$$L_s(\theta \mid X_n, Y_n) = \exp\{l_s(\theta)\} \cong \exp\left\{l_s(\hat{\theta}) + \dot{l}_s(\hat{\theta})(\theta - \hat{\theta}) - \frac{1}{2}(\theta - \hat{\theta})^{\mathrm{T}} I_s(\hat{\theta})(\theta - \hat{\theta})\right\},$$

where $\dot{l}_s(\theta) = \partial l_s(\theta)/\partial\theta^{\mathrm{T}}$, $I_s(\theta) = -\partial^2 l_s(\theta)/(\partial\theta\partial\theta^{\mathrm{T}})$, and $A_n \cong B_n$ means that $A_n - B_n = o_{\mathrm{p}}(1)$. Since $\hat{\theta}$ satisfies $\dot{l}_s(\hat{\theta}) = 0$, we have $L_s(\theta \mid X_n, Y_n) \cong \exp\{l_s(\hat{\theta})\} \times \exp\{-(\theta - \hat{\theta})^{\mathrm{T}} I_s(\hat{\theta})(\theta - \hat{\theta})/2\}$. Thus, writing $g(\hat{\theta} \mid \theta) \propto \exp\{-(\theta - \hat{\theta})^{\mathrm{T}} I_s(\hat{\theta})(\theta - \hat{\theta})/2\}$, we have $L_s(\theta \mid X_n, Y_n) \cong g(\hat{\theta} \mid \theta)K(X_n, Y_n)$ for some $K(X_n, Y_n)$ that does not depend on θ, which proves the equivalence between (2) and (4). The statistic $\hat{\theta} = \hat{\theta}(X_n, Y_n)$ is essentially a sufficient statistic for θ. In general, $g(\hat{\theta} \mid \theta)$ can be well approximated by a normal distribution with mean θ and variance $\mathrm{var}(\hat{\theta} \mid \theta)$ for large samples.

Remark 2. The proposed multiple imputation procedure extends readily to the case with multivariate variables that are subject to item nonresponse. Let $Y_{\text{obs},i}$ and $Y_{\text{mis},i}$ be the observed and missing items for unit i. In the new procedure, the P-step remains the same, which uses the posterior density of θ given $\hat{\theta}$. The I-step now generates $Y_{\text{mis},i}^* \sim f(Y_{\text{mis},i} \mid Y_{\text{obs},i}; \theta^*)$, which can be obtained by Bayes formula according to different missing patterns. If $f(Y_{\text{mis},i} \mid Y_{\text{obs},i}; \theta)$ is not in a closed form, Monte Carlo methods are needed in this step.

3. MAIN RESULT

To discuss the asymptotic properties of our procedure, we first assume a sequence of finite populations and samples with finite fourth moments as in Isaki & Fuller (1982). The finite population is a random sample from a superpopulation model ζ as presented in § 2. We assume the following regularity conditions:

Condition 1. Sufficient conditions for asymptotic normality of the pseudo maximum likelihood estimator hold for the sequence of finite populations and samples.

Condition 2. The prior density π is positive and satisfies a Lipschitz condition over Θ, i.e., there exists $C_1 < \infty$ such that $|\pi(\theta_1) - \pi(\theta_2)| \leqslant C_1 \|\theta_1 - \theta_2\|$.

Condition 3. Let B_n be the ball of centre θ_0 with radius $r_n \sim n^{\tau-1/2}$ for $0 < \tau < 1/2$. For any $\theta \in B_n$, the variance estimator $\hat{V}(\hat{\theta})$ satisfies $\mathrm{var}(\hat{\theta} \mid \theta) = \hat{V}(\hat{\theta})\{1 + o_{\mathrm{p}}(1)\}$ and $(\hat{\theta} - \theta)^{\mathrm{T}}\mathrm{var}(\hat{\theta} \mid \theta)^{-1}(\hat{\theta} - \theta) = (\hat{\theta} - \theta)^{\mathrm{T}}\hat{V}(\hat{\theta})^{-1}(\hat{\theta} - \theta)\{1 + o_{\mathrm{p}}(1)\}$ as $n \to \infty$.

Sufficient conditions for Condition 1 are discussed, for example, in Chapter 1 of Fuller (2009). Condition 2 is satisfied for classical prior densities, e.g., a flat prior over a bounded domain. Condition 3 is not straightforward. For illustration, we discuss a set of sufficient conditions under simple random sampling with $\mathrm{var}(\hat{\theta} \mid \theta) = n^{-1}I(\theta)^{-1}$ and $\hat{V}(\hat{\theta}) = n^{-1}I(\hat{\theta})^{-1}$, where $I(\theta)$ is the Fisher information matrix. Assume $I(\theta)$ and $x^{\mathrm{T}}I(\theta)x$ satisfy the following Lipschitz conditions over Θ. There exists $C_2 < \infty$ such that for any $\theta_1, \theta_2 \in \Theta$, $\||I(\theta_1)| - |I(\theta_2)|\| \leqslant C_2\|\theta_1 - \theta_2\|$. For any x, there exists $C_3(x) < \infty$ such that for any $\theta_1, \theta_2 \in \Theta$, $|x^{\mathrm{T}}I(\theta_1)x - x^{\mathrm{T}}I(\theta_2)x| \leqslant C_3(x)\|\theta_1 - \theta_2\|$. We further assume that there exists $C_4 < \infty$ such that for any x_1 and x_2, $|C_3(x_1) - C_3(x_2)| \leqslant C_4\|x_1 - x_2\|$. Thus, $|(\hat{\theta} - \theta)^{\mathrm{T}}\mathrm{var}(\hat{\theta} \mid \theta)^{-1}(\hat{\theta} - \theta) - (\hat{\theta} - \theta)^{\mathrm{T}}\hat{V}(\hat{\theta})^{-1}(\hat{\theta} - \theta)|$ is bounded by $O(r_n/n)$ for any $\theta \in B_n$, implying Condition 3. Sketch proofs of Lemma 1 and Theorem 1 are given in the Supplementary Material.

LEMMA 1. *Under Condtions* 1–3*, conditional on the full sample data,*

$$p_g(\theta \mid X_n, Y_n) \to N\{\hat{\theta}, \hat{V}(\hat{\theta})\} \tag{7}$$

in distribution as $n \to \infty$ almost surely.

The asymptotic result in (7) can be called the Bernstein–von Mises theorem (van der Vaart, 1998, Ch. 10) for the posterior density (4) induced by the sampling distribution $g(\hat{\theta} \mid \theta)$. By Lemma 1,

$$E_g(\theta \mid X_n, Y_n) = \hat{\theta}\{1 + o_{\mathrm{p}}(1)\}, \quad \mathrm{var}_g(\theta \mid X_n, Y_n) = \hat{V}(\hat{\theta})\{1 + o_{\mathrm{p}}(1)\} \tag{8}$$

almost surely, where the reference distribution is the posterior density (4).

THEOREM 1. *Under Conditions* 1–3 *and the population missingness-at-random assumption, ignoring smaller-order terms,*

$$p\lim_{M\to\infty} \hat{\theta}_{\mathrm{MI}} = E_g(\theta \mid X_n, Y_{\mathrm{obs}}), \quad p\lim_{M\to\infty} \hat{V}_{\mathrm{MI}} = \mathrm{var}_g(\theta \mid X_n, Y_{\mathrm{obs}}),$$

where the conditional expectations are with respect to posterior density (6).

By Lemma 1 and Theorem 1, $\hat{V}_{\mathrm{MI}}^{-1/2}(\hat{\theta}_{\mathrm{MI}} - \theta) \to N(0, 1)$ in distribution as $n \to \infty$ and $M \to \infty$.

We now discuss inference for an induced parameter $\gamma = \gamma(\theta)$, such as $\gamma = \int_{-\infty}^{1} f(y; \theta)\,\mathrm{d}y$. In this case, the posterior distribution of $\gamma = \gamma(\theta)$ is directly obtained from the distribution of $\gamma(\theta^*)$, where θ^* is generated from $p_g(\theta \mid X_n, Y_{\mathrm{obs}})$ in (6). Under certain regularity conditions, the posterior density of γ, $p_g(\gamma \mid X_n, Y_{\mathrm{obs}})$, converges to the normal distribution with mean $\gamma(\hat{\theta})$ and variance $\gamma'(\hat{\theta})^2\hat{V}(\hat{\theta})$ almost surely by the continuous mapping theorem (Mann & Wald, 1943). If $\gamma'(\theta)$ is bounded, it is straightforward to show that the corresponding properties in Theorem 1 hold for $\hat{\gamma}_{\mathrm{MI}}$ and the corresponding variance estimator.

4. SIMULATION STUDY

In this simulation study, we assess the finite-sample performance of the proposed multiple imputation procedure. We consider the outcome to be continuous or binary, combined with nonignorable or ignorable sampling.

In the first set-up, we consider a continuous outcome which follows the following superpopulation model, $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, where $x_i \sim N(0, 1)$, $\epsilon_i \sim N(0, \sigma^2)$, and $\theta = (\beta_0, \beta_1, \sigma^2) = (-0{\cdot}5, 0{\cdot}5, 1)$.

Table 1. *Monte Carlo biases and standard errors of the point estimators of* $\eta = N^{-1}\sum_{i=1}^{N} y_i$*, along with the Monte Carlo coverage* (%) *of* 95% *confidence intervals based on* 5000 *simulated samples*

	Setting		I. Ignorable			II. Nonignorable	
Model	Method	Bias	SE	Coverage	Bias	SE	Coverage
		($\times 10^2$)	($\times 10^2$)	(%)	($\times 10^2$)	($\times 10^2$)	(%)
Linear Model	Hajek	0	40	95	0	33	95
	Traditional MI	0	45	95	4	38	83
	Proposed MI	0	45	95	0	40	95
Logistic Model	Hajek	0	18	95	0	12	95
	Traditional MI	0	20	95	3	14	41
	Proposed MI	0	19	95	0	13	95

MI, multiple imputation with imputation size 100; SE, standard error.

We first generate finite populations of size $N = 50000$. The response indicator of y_i is generated from $\delta_i \sim \text{Ber}(\phi_i)$, where

$$\text{logit}\,\phi_i = 1 + 0{\cdot}5x_i + 0{\cdot}5u_i, \tag{9}$$

with $u_i \sim N(0,1)$, and u_i is independent of x_i and ϵ_i. By construction, $y \perp\!\!\!\perp \delta \mid x$, i.e., missingness at random holds at the population level. For the sampling mechanism, we use Poisson sampling with $I_i \sim \text{Ber}(\pi_i)$. We consider two settings: Setting I with $\text{logit}(1-\pi_i) = 4 + 0{\cdot}5x_i$, and Setting II with $\text{logit}(1-\pi_i) = 3{\cdot}66 + 0{\cdot}33u_i - 0{\cdot}1y_i$. Since in Setting II, u_i is a common cause for δ_i and I_i, and y_i is a cause for I_i, we have $y_i \not\perp\!\!\!\perp \delta_i \mid (x_i, w_i, I_i = 1)$, i.e., the missingness mechanism becomes nonignorable at the sample level under the augmented model.

In the second set-up, we consider a binary outcome which follows a logistic regression superpopulation model, $y_i \sim \text{Bin}(p_i)$, where $p_i = \exp(\beta_0 + \beta_1 x_i)/\{1 + \exp(\beta_0 + \beta_1 x_i)\}$, $x_i \sim N(0,1)$, and $\theta = (\beta_0, \beta_1) = (-0{\cdot}5, 0{\cdot}5)$. The response indicator of y_i is generated in the same way according to (9). For the sampling mechanism, in Setting I, $\text{logit}(1-\pi_i) = 4 + 0{\cdot}5x_i$, and in Setting II, $\text{logit}(1-\pi_i) = 3{\cdot}66 + 0{\cdot}33u_i - 0{\cdot}5y_i$. The average sample sizes range from 1000 to 1600 and the average response rates are around 65%. In the Supplementary Material, we also investigate scenarios where the average sample sizes range from 160 to 200. The results are similar.

We consider estimating η, the population mean of y, with the following estimators: (i) the Hajek estimator, applied to the full sample, assuming all observations are available, which serves as a benchmark for comparison; (ii) the traditional multiple imputation estimator; and (iii) the proposed multiple imputation estimator. For (ii), for the continuous outcome, we assume the imputation model $f(y \mid x, w; \theta)$ to be a linear regression model of y on x and w, where $w = 1/\pi$ is the design weight. For the binary outcome, we assume the imputation model $f(y \mid x, w; \theta)$ to be a logistic regression model. For (iii), we use the correctly specified superpopulation model, and the sampling distribution $g(\hat\theta \mid \theta)$ in (5) is a normal distribution, with mean θ and variance $V(\hat\theta)$. Here, $\hat\theta$ is the pseudo maximum likelihood estimator of θ, which is obtained by solving (3). For the design-consistent estimator of $V(\hat\theta)$, we use $\hat V(\hat\theta) = \hat\tau^{-1}\hat V(S)(\hat\tau^{-1})^{\mathrm{T}}$, where $\hat\tau = \sum_{i\in A} w_i \dot S(\hat\theta; x_i, y_i)$, $\dot S(\theta; x, y) = \partial S(\theta; x, y)/\partial\theta^{\mathrm{T}}$, and $\hat V(S) = \sum_{i\in A}(w_i^2 - w_i)\hat S_i \hat S_i^{\mathrm{T}}$, with $\hat S_i = S(\hat\theta; x_i, y_i)$. For both (i) and (ii), the priors for regression coefficients are independent normal distributions with mean 0 and variance 10^6, the prior for σ^2 is uniform over the interval $[0, 10^6]$, and the complete-sample estimator of the population mean of y, for each imputed dataset, is $\hat\eta = (\sum_{i\in A} w_i)^{-1} \sum_{i\in A} w_i y_i$, while the variance estimator of $\hat\eta$ is $\hat V(\hat\eta) = (\sum_{i\in A} w_i)^{-2} \times \sum_{i\in A}(w_i^2 - w_i)(y_i - \hat\eta)^2$.

Table 1 shows the simulation results over 5000 Monte Carlo samples. In Setting I, both multiple imputation methods provide valid inference because the sampling design is ignorable and the sample-data likelihood function can be derived directly from the density function in the superpopulation model. However, in Setting II, the traditional method is biased, because $y \perp\!\!\!\perp \delta \mid (w, x, I = 1)$ is violated in Setting II.

As a result, in Setting II, the coverage of the confidence interval for the traditional method is quite poor. The proposed method is essentially unbiased and has good coverage, confirming our theoretical results. It does not lose efficiency compared to the traditional method in Setting I under the ignorable sampling design, even though the parameters are generated by an approximated Bayesian computation approach, because the complete-sample estimator of η is a design-weighted estimator, which is not necessarily self-efficient (Meng, 1994). In this case, using a more efficient approach to generating parameters may not result in a more efficient multiple imputation estimator of η.

Acknowledgement

We are grateful to Professor Donald B. Rubin, the editor and three referees for their valuable comments that have helped to improve this paper. The research of the first author was partially supported by the U.S. National Science Foundation. J. K. Kim is also affiliated with the Korean Advanced Institute of Science & Technology.

Supplementary material

Supplementary material available at *Biometrika* online includes the proofs of Lemma 1 and Theorem 1, as well as additional simulation results.

References

Berg, E., Kim, J. K. & Skinner, C. (2016). Imputation under informative sampling. *J. Survey Statist. Methodol.* **8**, 1–27.

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *Int. Statist. Rev.* **51**, 279–92.

Chambers, R. L. & Skinner, C. J. (2003). *Analysis of Survey Data*. New York: Wiley.

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

Fearnhead, P. & Prangle, D. (2012). Constructing summary statistics for approximate Bayesian computation: Semi-automatic approximate Bayesian computation (with Discussion). *J. R. Statist. Soc.* B **74**, 419–74.

Fuller, W. A. (2009). *Sampling Statistics*. Hoboken: Wiley, 3rd ed.

Godambe, V. & Thompson, M. E. (1986). Parameters of superpopulation and survey population: Their relationships and estimation. *Int. Statist. Rev.* **54**, 127–18.

Isaki, C. T. & Fuller, W. A. (1982). Survey design under the regression superpopulation model. *J. Am. Statist. Assoc.* **77**, 89–96.

Kim, J. K., Brick, J., Fuller, W. A. & Kalton, G. (2006). On the bias of the multiple-imputation variance estimator in survey sampling. *J. R. Statist. Soc.* B **68**, 509–21.

Kim, J. K. & Park, M. (2010). Calibration estimation in survey sampling. *Int. Statist. Rev.* **78**, 21–39.

Korn, E. L. & Graubard, B. I. (2011). *Analysis of Health Surveys*. New York: Wiley.

Kott, P. (1995). A paradox of multiple imputation. In *Proc. Survey Res. Meth. Sect.* American Statistical Association, pp. 380–3.

Little, R. J. (2012). Calibrated Bayes, an alternative inferential paradigm for official statistics. *J. Offic. Statist.* **28**, 309–34.

Little, R. J. & Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. Hoboken: Wiley, 2nd ed.

Mann, H. B. & Wald, A. (1943). On stochastic limit and order relationships. *Ann. Math. Statist.* **14**, 217–26.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statist. Sci.* **9**, 538–58.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *Int. Statist. Rev.* **61**, 317–37.

Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H. & Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *J. R. Statist. Soc.* B **60**, 23–40.

Pfeffermann, D. & Sverchkov, M. (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhyā* B **61**, 166–86.

Reiter, J. P., Raghunathan, T. E. & Kinney, S. K. (2006). The importance of modeling the sampling design in multiple imputation for missing data. *Survey Methodol.* **32**, 143–50.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581–92.

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

Rubin, D. B. (1996). Multiple imputation after 18+ years. *J. Am. Statist. Assoc.* **91**, 473–89.

Rubin, D. B. & Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *J. Am. Statist. Assoc.* **81**, 366–74.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. London: CRC Press.

Schenker, N., Raghunathan, T. E., Chiu, P.-L., Makuc, D. M., Zhang, G. & Cohen, A. J. (2006). Multiple imputation of missing income data in the National Health Interview Survey. *J. Am. Statist. Assoc.* **101**, 924–33.

Schenker, N. & Welsh, A. (1988). Asymptotic results for multiple imputation. *Ann. Statist.* **16**, 1550–66.

Scott, A. (1977). On the problem of randomization in survey sampling. *Sankhyā* C **39**, 1–9.

Soubeyrand, S. & Haon-Lasportes, E. (2015). Weak convergence of posteriors conditional on maximum pseudo-likelihood estimates and implications in ABC. *Statist. Prob. Lett.* **107**, 84–92.

Tanner, M. A. & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *J. Am. Statist. Assoc.* **82**, 528–40.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge: Cambridge University Press.

Wang, N. & Robins, J. M. (1998). Large-sample theory for parametric multiple imputation procedures. *Biometrika* **85**, 935–48.

Yang, S. & Kim, J. K. (2016). A note on multiple imputation for method of moments estimation. *Biometrika* **103**, 244–51.

Yuan, Y. & Little, R. J. (2007). Model-based estimates of the finite population mean for two-stage cluster samples with unit non-response. *J. R. Statist. Soc.* C **56**, 79–97.

[*Received on* 29 *February* 2016. *Editorial decision on* 5 *November* 2016]