On the bias of the multiple-imputation variance estimator in survey sampling

Jae Kwang Kim,
*Yonsei University, Seoul, Korea*

J. Michael Brick,
*Westat, Rockville, USA*

Wayne A. Fuller
*Iowa State University, Ames, USA*

and Graham Kalton
*Westat, Rockville, USA*

[Received September 2003. Final revision November 2005]

**Summary.** Multiple imputation is a method of estimating the variances of estimators that are constructed with some imputed data. We give an expression for the bias of the multiple-imputation variance estimator for data that are collected with a complex sample design. The bias may be sizable for certain estimators, such as domain means, when a large fraction of the values are imputed. A bias-adjusted variance estimator is suggested.

**Keywords:** Missing data; Non-response; Survey sampling

1. Introduction

Imputation is widely used in sample surveys to assign values for item non-responses. If the imputed values are treated as if they were observed then estimates of the variances of the estimates will generally be underestimates. Multiple imputation (MI) has been proposed as a method for estimating the precision of sample estimates in the presence of imputed values (see Rubin (1987, 1996)). MI is applied to a data set with missing items by repeating the process of assigning values for each of the missing values \( M \) times, creating \( M \) completed data sets. Each of the \( M \) completed data sets can be used to compute an estimate \( \hat{\theta}_{I(k)} \) of a population parameter \( \theta \) \((k = 1, 2, \ldots, M)\), with the subscript \( I \) indicating that some values have been imputed.

Rubin (1987) proposed that \( \theta \) be estimated by the average of the \( M \) estimates

\[
\hat{\theta}_{M,n} = M^{-1} \sum_{k=1}^{M} \hat{\theta}_{I(k)},
\]

and that the variance of \( \hat{\theta}_{M,n} \) be estimated by

\[
\hat{V}_{M,n} = U_{M,n} + (1 + M^{-1}) B_{M,n},
\]

**Address for correspondence:** J. Michael Brick, Westat, 1650 Research Boulevard, Rockville, MD 20850, USA.
E-mail: mikebrick@westat.com

© 2006 Royal Statistical Society 1369–7412/06/68509
where \( n \) is the original sample size,

\[
U_{M,n} = M^{-1} \sum_{k=1}^{M} \hat{V}_{I(k)},
\]

\[
B_{M,n} = (M - 1)^{-1} \sum_{k=1}^{M} (\hat{\theta}_{I(k)} - \hat{\theta}_{M,n})^2
\]

and \( \hat{V}_{I(k)} \) is the variance estimator computed from the \( k \)th data set treating imputed values as if they were observed values. The estimators contain a subscript for the sample size \( n \) because asymptotic results are functions of \( M \) and \( n \). We call equation (1.2) the MI variance estimator.

This paper investigates the bias of the MI variance estimator in the sample survey context. Survey samples are generally based on complex sample designs, and weights that are based on the inverse of selection probabilities are often adjusted to compensate for unit non-response and to make sample estimates conform to benchmark totals. This paper examines estimators that can be written as

\[
\hat{\theta}_n = \sum_{i \in A} \alpha_i Y_i
\]

for known coefficients \( \alpha_i \), where \( A \) is the set of indices of the sample elements. The estimation weights \( \alpha_i \) can include non-response and calibration adjustments but cannot be functions of the \( Y_i \). Domain estimates are obtained by setting \( \alpha_i = 0 \) for sample elements that are outside the domain.

Given an appropriate imputation scheme, the MI variance estimator is intended to be applicable to any estimator \( \hat{\theta}_{M,n} \). Rubin (1987) discussed MI both from a Bayesian perspective and from a randomization or quasi-randomization perspective. In the latter case the reference distribution is the joint distribution determined by the sampling distribution and by the response mechanism. The problem of MI under the randomization approach is that the validity of the MI variance estimator requires the assumption of proper imputation as defined by Rubin (1987), and no method for constructing a proper imputation procedure has been given for complex sample surveys. Binder and Sun (1996) investigated conditions that are required for proper imputation for estimates of means and totals from general complex sampling schemes and concluded that the conditions of proper imputation are very difficult to satisfy. Fay (1992, 1993) demonstrated that proper imputation for the mean may not be proper imputation for a domain mean, when the domain indicator is not included in the imputation model.

Under a parametric model approach, Wang and Robins (1998) and Nielsen (2003) showed that the MI variance estimator for \( M = \infty \) may be biased unless \( \hat{\theta}_n \) is the maximum likelihood estimator of the parameter \( \theta \) in the imputer’s parametric model. Robins and Wang (2000) extended the result to nonparametric models and proposed an alternative variance estimator. However, their results are not directly applicable to complex sample surveys.

This paper considers the applicability of the MI variance estimator for general complex survey sampling schemes under a superpopulation model in which the finite survey population is assumed to be a random sample from an infinite population. The superpopulation model approach has been used in the application of MI to sample surveys by, for example, Clogg et al. (1991), Schafer et al. (1996), Gelman et al. (1998), Davey et al. (2001) and Taylor et al. (2002). Complex survey sampling schemes involve weights in their analyses. Even when a superpopulation model is assumed, it is common practice to use survey weights as a protection against model misspecification (e.g. Pfeffermann (1993)).
Kott (1995) investigated the bias of the MI variance estimator for the weighted sample mean under a simple mean superpopulation model and showed that the MI variance estimator may be biased unless the weights are incorporated into the imputation process. We extend Kott’s results to more general superpopulation models and to a more general class of point estimators. The basic assumptions are introduced in Section 2. We examine the bias of $\hat{V}_{M,n}$ as an estimator for $\text{var}(\hat{\theta}_{M,n})$ in Section 3. We apply the results of Section 3 to variance estimation for domain estimators in Section 4. Extension to a multiple-regression model is discussed in Section 5.

2. Basic set-up

To study the properties of the MI variance estimator, we consider the joint distribution that is defined by the superpopulation model, the sampling mechanism, the response mechanism and the imputation mechanism. Unconditional expectations, variances and covariances with respect to all these factors are denoted by $E(\cdot)$, $\text{var}(\cdot)$ and $\text{cov}(\cdot)$. We derive the properties of the MI variance estimator under the following assumptions.

(a) Condition (C.1): the complete-sample point estimator $\hat{\theta}_n$ is of the form given in equation (1.3) and satisfies

$$E(\hat{\theta}_n) = \theta + O(n^{-1}).$$

The complete-sample variance estimator $\hat{V}_n$ is quadratic in the $y$-variable:

$$\hat{V}_n = \sum_{i \in A} \sum_{j \in A} \Omega_{ij} Y_i Y_j$$

for some coefficients $\Omega_{ij}$ and

$$E(\hat{V}_n) = \text{var}(\hat{\theta}_n) + o(n^{-1}).$$

(b) Condition (C.2): the sampling mechanism and the response mechanism are ignorable under the superpopulation model in the sense of Rubin (1976). Thus, for any $\sigma(Y)$ measurable sets $B_1$ and $B_2$,

$$\Pr(Y_i \in B_1, Y_j \in B_2) = \Pr(Y_i \in B_1, Y_j \in B_2 \mid A, A_R),$$

where $A_R$ is the set of indices of the responding sample elements.

(c) Condition (C.3): conditional on $A$ and $A_R$, the imputed and real values have the same expected values

$$E(\eta_{i(k)} \mid A, A_R) = E(Y_i \mid A, A_R),$$

where $\eta_{i(k)}$ is the $k$th imputed value for unit $i$. If $Y_i$ is observed, $\eta_{i(k)} = Y_i$ for all $k$.

(d) Condition (C.4): conditional on $A$ and $A_R$, the imputed and real values have the same asymptotic covariances

$$\max_{i,j} |\text{cov}(\eta_{i(k)}, \eta_{j(k)} \mid A, A_R) - \text{cov}(Y_i, Y_j \mid A, A_R)| = o(1),$$

for all $k$.

(e) Condition (C.5): for each $i \in A_M$, where $A_M$ is the set of indices for the non-respondents, the $M$ values of $\eta_{i(k)}$ are independently and identically distributed, conditional on $A, A_R$, and $y_r = \{Y_i; i \in A_R\}$.

Assumption (C.2) is the standard condition that is used with imputation under the superpopulation model (see, for example, Särndal (1992)). Under assumption (C.2), the order of the
expectation operator can be changed. Under assumptions (C.1) and (C.3), all imputed estimators of the form (1.3) are approximately unbiased. Under assumption (C.4), the covariances of the imputed values mimic those of the original values. An important reason for using a stochastic imputation scheme is to accomplish condition (C.4). Many superpopulation models assume that \( \text{cov}(Y_i, Y_j) = 0 \) for \( i \neq j \), in which case condition (C.4) together with condition (C.2) imply that \( \text{cov}(\eta_i, \eta_j|A, A_R) = o(1) \) for \( i \neq j \) and \( \text{var}(\eta_i|A, A_R) = \text{var}(Y_i|A, A_R) + o(1) \). When \( \text{cov}(Y_i, Y_j) = 0 \) for \( i \neq j \), assumption (C.4) is satisfied by some stochastic imputation methods, including the approximate Bayesian bootstrap imputation of Rubin and Schenker (1986) and the regression imputation that is described in Section 4. Under assumption (C.5), the \( M \) point estimators \( \hat{\theta}_{l(k)} \) are identically distributed and the \( M \) naïve variance estimators \( \hat{V}_{l(k)} \) are identically distributed.

3. Main results

To derive the expected value of the MI variance estimator, we first express \( \hat{\theta}_{M,n} \) as

\[
\hat{\theta}_{M,n} = \hat{\theta}_n + (\hat{\theta}_{\infty,n} - \hat{\theta}_n) + (\hat{\theta}_{M,n} - \hat{\theta}_{\infty,n})
\]

where \( \hat{\theta}_{\infty,n} = \lim_{M \to \infty} (\hat{\theta}_{M,n}) \) is the infinite \( M \) MI point estimator. We call the variance of the first term on the right-hand side of equation (3.1) the sampling variance, the variance of the second term the variance due to missingness and the variance of the third term the imputation variance. In the MI variance estimator, \( U_{M,n} \) is an estimator of the sampling variance, \( B_{M,n} \) is an estimator of the variance due to missingness and \( M^{-1}B_{M,n} \) is an estimator of the imputation variance. If the three terms in equation (3.1) are uncorrelated, then the MI variance estimator will be approximately unbiased. However, it is possible for \( \hat{\theta}_n \) to be correlated with the second term. We show that the bias of the MI variance estimator is a function of the covariance between the full sample estimator and the MI point estimator, a result which is consistent with Kott’s (1995) conclusion, which was derived in a simpler setting.

Since \( U_{M,n} = M^{-1} \sum_{k=1}^{M} \hat{V}_{l(k)} \) and the \( \hat{V}_{l(k)} \) are identically distributed under assumption (C.5), examining the bias of \( U_{M,n} \) is equivalent to examining the bias of the naïve variance estimator \( \hat{V}_{l(k)} \) for any \( k \). Thus, given conditions (C.1)–(C.3),

\[
E(\hat{V}_l) - \text{var}(\hat{\theta}_n) = E\left( \sum_{i \in A} \sum_{j \in A} \Omega_{ij} \tau_{ij} \right) + o(n^{-1})
\]

where the \( \Omega_{ij} \) are the coefficients in equation (2.1) and

\[
\tau_{ij} = \text{cov}(\eta_i, \eta_j|A, A_R) - \text{cov}(Y_i, Y_j|A, A_R)
\]

is the difference between the covariance of the imputed values and the covariance of the original values. Note that \( \tau_{ij} = 0 \) when both \( Y_i \) and \( Y_j \) are observed.

An estimated full sample variance that is consistent and nearly unbiased exists for many sampling designs, including stratified cluster sampling designs. Consider an estimator of a sample mean with \( \sum_{i \in A} \alpha_i = 1 \) and assume a sequence of finite populations with finite fourth moments as described in Isaki and Fuller (1982). Let the sequence of estimators satisfy

\[
\sum_{i \in A} \sum_{j \in A} |\Omega_{ij}| = O(n^{-1}).
\]

Then, by equation (3.2) and conditions (C.4) and (C.5),

\[
E(\hat{U}_{M,n}) = \text{var}(\hat{\theta}_n) + o(n^{-1}).
\]
The following lemma gives the bias of $B_{M,n}$ as an estimator of the variance due to missingness.

**Lemma 1.** Let the complete-sample point estimator be of the form (1.3) and let conditions (C.1)–(C.3) and (C.5) hold. Then,

$$E(B_{M,n}|A, A_R) - \text{var}(\hat{\theta}_{\infty,n} - \hat{\theta}_n|A, A_R) = \text{var}\left(\sum_{i \in A_M} \alpha_i \eta_i(1)|A, A_R\right) - \text{var}\left(\sum_{i \in A_M} \alpha_i Y_i|A, A_R\right)$$

$$- 2 \text{cov}\left(\sum_{i \in A_M} \alpha_i \eta_i(1), \sum_{i \in A_M} \alpha_i \eta_i(2)|A, A_R\right),$$

where $A_M$ is the set of indices for the non-respondents.

Furthermore, if we also assume condition (C.4), assume $\text{cov}(Y_i, Y_j) = 0$ for $i \neq j$ and assume

$$\text{cov}(\eta_i(1), \eta_j(1)|A, A_R) = 2 \text{cov}(\eta_i(1), \eta_j(2)|A, A_R) + o(n^{-1})$$

for $i \neq j$ and $i, j \in A_M$, then

$$E(B_{M,n}|A, A_R) - \text{var}(\hat{\theta}_{\infty,n} - \hat{\theta}_n|A, A_R) = o(n^{-1}).$$

For a proof, see Appendix A.

Condition (3.6) is satisfied with MI when different stochastically generated parameter estimates are used in constructing imputed values in different imputed data sets, as is done in some Bayesian bootstrap imputation procedures. For example, this condition holds with the approximate Bayesian bootstrap imputation procedure that was proposed by Rubin and Schenker (1986) and with the regression imputation procedure of Schenker and Welsh (1988).

The following lemma shows that $M^{-1}B_{M,n}$ is an unbiased estimator of the imputation variance.

**Lemma 2.** Given condition (C.5),

$$E(M^{-1}B_{M,n}|A, A_R) = \text{var}(\hat{\theta}_{M,n} - \hat{\theta}_{\infty,n}|A, A_R)$$

and

$$\text{cov}(\hat{\theta}_{\infty,n} - \hat{\theta}_n, \hat{\theta}_{M,n} - \hat{\theta}_{\infty,n}|A, A_R) = 0$$

for all $M \geq 2$ and all $n$.

For a proof, see Appendix A.

By equation (3.4), $U_{M,n}$ will be nearly unbiased for the sampling variance when the imputation procedure reproduces the basic covariance structure of the original sample. By lemma 1, $B_{M,n}$ is nearly unbiased for the variance due to missingness when the imputation procedure makes an allowance for the variability due to parameter estimation. By lemma 2, the component $M^{-1}B_{M,n}$ will be an unbiased estimator of the imputation variance for any MI scheme satisfying condition (C.5). Therefore, the primary potential source of bias in the MI variance estimator is the fact that the error in $\hat{\theta}_{M,n}$, as an estimator of $\hat{\theta}_n$, may be correlated with $\hat{\theta}_n$.

**Theorem 1.** Assume a sequence of finite populations selected from a superpopulation with $E(Y_i^4) < \infty$. Let conditions (C.1)–(C.5) hold. Let the sequence of complete-sample point estimators of the form (1.3) have variance of order $n^{-1}$ with $\max_i(\alpha_i) = O(n^{-1})$. Let the complete-sample variance estimator satisfy condition (3.3). Assume that the MI procedure is such that

$$E(B_{M,n}|A, A_R) = \text{var}(\hat{\theta}_{\infty,n} - \hat{\theta}_n|A, A_R) + o(n^{-1}).$$

(3.10)
Then, \[
E(\hat{V}_{M,n}) - \text{var}(\hat{\theta}_{M,n}) = -2 E\{\text{cov}(\hat{\theta}_{I(1)} - \hat{\theta}_n, \hat{\theta}_n | A, A_R)\} + o(n^{-1}) \tag{3.11}
\]
for all \(M \geq 2\).

For a proof, see Appendix A.

Kott (1995) showed that the bias of the MI variance estimator for the simple mean model can be expressed as a covariance. Theorem 1 is an extension to more general estimators. Särndal (1992) considered variance estimation for imputed samples with single imputation using a decomposition that was similar to equation (3.1). He expressed the total variance as
\[
V_{\text{TOT}} = V_{\text{SAM}} + V_{\text{IMP}} + 2V_{\text{MIX}},
\]
where \(V_{\text{SAM}} = E(\hat{\theta}_n - \theta)^2\), \(V_{\text{IMP}} = E(\hat{\theta}_I - \hat{\theta}_n)^2\) and \(V_{\text{MIX}} = E\{(\hat{\theta}_n - \theta)(\hat{\theta}_I - \hat{\theta}_n)\}\). Särndal’s \(V_{\text{MIX}}\) is the covariance that is given in equation (3.11). Meng (1994) gave sufficient conditions for the MI variance estimator to be unbiased. One condition, called ‘self-efficient’, can be expressed in our notation as
\[
\text{var}(\hat{\theta}_{\infty,n}) = \text{var}(\hat{\theta}_n) + \text{var}(\hat{\theta}_{\infty,n} - \hat{\theta}_n). \tag{3.12}
\]
Note that condition (3.12) is equivalent to assuming that the bias of equation (3.11) is 0.

To investigate the implications of equation (3.11), assume that the expected value of the imputed value for recipient \(j\) can be expressed as a linear function of the observed sample values,
\[
E(\eta_{j(1)}|y_r, A, A_R) = \sum_{i \in A_R} \gamma_{ji} Y_i, \tag{3.13}
\]
where \(y_r\) is the vector of respondent values and \(\gamma_{ji}\) are fixed coefficients. Examples of situations where the expectation has this form include the use of imputation cells with donors which are selected by the approximate Bayesian bootstrap, normal distribution imputation and regression imputation as described by Rubin and Schenker (1986). Then, for an estimator of the form (1.3),
\[
\text{cov}(\hat{\theta}_{I(1)} - \hat{\theta}_n, \hat{\theta}_n | A, A_R) = \text{cov}\left\{ \sum_{j \in A_M} \alpha_j \left( \sum_{i \in A_R} \gamma_{ji} Y_i - Y_j \right), \sum_{i \in A} \alpha_i Y_i | A, A_R \right\}.
\]
Assume that, under the model, \(Y_i\) is independent of \(Y_j, i \neq j\), and assume a common variance \(\sigma^2\). Then
\[
\text{cov}(\hat{\theta}_{I(1)} - \hat{\theta}_n, \hat{\theta}_n | A, A_R) = \left( \sum_{j \in A_M} \alpha_j \sum_{i \in A_R} \gamma_{ji} \alpha_i - \sum_{j \in A_M} \alpha_j^2 \right) \sigma^2
\]
and
\[
\text{bias}(\hat{V}_{M,n}) = 2 E\left\{ \sum_{j \in A_M} \alpha_j (\alpha_j - \hat{\alpha}_{Ij}) \right\} \sigma^2 \tag{3.14}
\]
where
\[
\hat{\alpha}_{Ij} = \sum_{i \in A_R} \gamma_{ji} \alpha_i.
\]
The \(\hat{\alpha}_{Ij}\) is the expected value of the ‘imputed value’ for \(\alpha_j\) that would be computed if \(\alpha_i\) had the missing pattern of the \(y\)s. Thus the nature of the bias in a variance that is estimated by MI can be determined by computing the imputed values for the \(\alpha\)-coefficients of the estimator that is associated with the missing \(y\)-values and then computing the summation in equation (3.14). The summation in equation (3.14) is the difference between the estimator \(\hat{\theta}\) in equation (1.3)
with \( Y_i = \alpha_i \) and the same estimator \( \hat{\theta} \) with \( Y_i = \hat{\alpha}_i \). Since \( \hat{\alpha}_ij = \alpha_j \) in \( A_R \), the summation for \( A_M \) is the same as the summation for \( A \).

### 4. Domain estimation

Estimation for subpopulations, called domains, is common in the analysis of survey data. Estimates are commonly presented in two-way or multiway tables, with the table cells defining the domains. If the imputation model does not contain an indicator for the domain, the model assumption for domain estimation is that the conditional distribution of \( Y \) in the domain, conditional on the model variables, is the same as the conditional distribution for the population. Under this assumption and retained assumptions (C.1)–(C.3), the imputed domain estimator is unbiased for the domain total.

To discuss the nature of the bias in the variance estimator for domain means, consider a simple random sample of size \( n \), with \( r \) respondents and \( m \) non-respondents. Assume that the finite population is a sample of independently and identically distributed random variables with mean \( \mu \) and variance \( \sigma^2 \). Assume that an imputation scheme such as the approximate Bayesian bootstrap of Rubin and Schenker (1986) is used so that assumptions (C.1)–(C.5) and (3.6) hold.

By equation (3.14), the MI variance estimator for the estimate of the overall population mean is asymptotically unbiased because \( \alpha_j = \hat{\alpha}_ij = n^{-1} \). Now consider estimation of a domain mean where the domain is defined by a variable \( z_i \), independent of \( y \), where \( z_i \) is 1 if element \( i \) is in domain \( D \) and is 0 otherwise. Assume that \( z \) is always observed. Let the full sample and imputed estimators of the domain mean be

\[
\hat{\theta}_{D,r} = \left( \sum_{i \in A} z_i \right)^{-1} \sum_{i \in A} z_i Y_i \quad (4.1)
\]

and

\[
\hat{\theta}_{l(k),D} = \left( \sum_{i \in A} z_i \right)^{-1} \sum_{i \in A} z_i \eta_{l(k)} \quad (4.2)
\]

respectively. Thus,

\[
\hat{\theta}_{\infty,D,r} = \hat{P}_D \bar{Y}_{r,D} + (1 - \hat{P}_D) \bar{Y}_r \quad (4.3)
\]

where

\[
\hat{P}_D = \left( \sum_{i \in A} z_i \right)^{-1} \sum_{i \in A_R} z_i
\]

is the observed response rate in domain \( D \),

\[
\bar{Y}_{r,D} = \left( \sum_{i \in A_R} z_i \right)^{-1} \sum_{i \in A_R} z_i y_i
\]

is the mean of respondents in domain \( D \) and \( \bar{Y}_r \) is the overall mean of respondents. Equation (4.3) is a weighted mean of the direct estimator \( \bar{Y}_{r,D} \) and the estimator \( \bar{Y}_r \) that is derived from the imputation model.

The conditional variance of the MI estimator of the domain mean is

\[
\text{var}(\hat{\theta}_{M,D,r} | A, A_R) = n_D^{-2} \{ r_D + m_D r^{-1} (m_D + 2r_D) + M^{-1} m_D^2 (m_D^{-1} + r^{-1}) \} \sigma^2,
\]

where

- \( M \) is the number of imputations,
- \( r_D \) is the response rate in domain \( D \),
- \( m_D \) is the number of non-respondents in domain \( D \),
- \( m_D^{-1} \) is the inclusion probability of a respondent in domain \( D \),
- \( M^{-1} \) is the inclusion probability of a unit in the overall sample,
- \( r^{-1} \) is the overall response rate.


where \( r_D \) is the number of respondents in domain \( D \), \( m_D \) is the number of non-respondents in domain \( D \), \( n_D = r_D + m_D \) and
\[
E\{B_{M,D,n}(\hat{\theta}_D)|A, A_R\} = n_D^{-2} m_D (1 + r^{-1} m_D) \sigma^2.
\]
Ignoring the \( M^{-1} \)-terms, the variance of the imputed estimator can be smaller than that of the complete-sample estimator if \( m_D + 2r_D < r \), which will happen for small domains. This is because, under the model, the best estimator for the domain mean is the mean of all respondents.

In this example \( \alpha_i \) is \( n_D^{-1} \) if element \( i \) is in domain \( D \) and is 0 otherwise. The expected value of an imputed \( \alpha_i \) as a function of the ‘respondent’ values is
\[
\hat{\alpha}_{ij} = r^{-1} \sum_{i \in A_R} \alpha_i
\]
and the conditional bias in the MI-estimated variance of the domain mean is
\[
2 E\left\{ \sum_{j \in A_M} \alpha_j (\alpha_j - \hat{\alpha}_{ij}) \sigma^2 | A, A_R \right\} = 2n_D^{-2} m_D (1 - r^{-1} r_D) \sigma^2.
\]
(4.4)
The conditional relative bias in the MI estimator of the domain variance is
\[
\frac{2 m_D r^{-1} (r - r_D)}{r_D + m_D r^{-1} (m_D + 2r_D) + M^{-1} m_D^2 (m_D^{-1} + r^{-1})}.
\]
Table 1 presents some numerical values of the conditional relative biases under various scenarios, assuming that the response rate in the domain is the same as that in the entire sample. The positive conditional relative biases in Table 1 are associated with the borrowing of strength from outside the domain due to imputation. The relative bias is larger for smaller domains and lower response rates. Even with a domain that comprises as much as 40% of the population and a

<table>
<thead>
<tr>
<th>( M )</th>
<th>Domain proportion (%)</th>
<th>% relative biases for the following response rates:</th>
</tr>
</thead>
<tbody>
<tr>
<td>( 5 )</td>
<td>( 10 )</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>( 20 )</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>( 30 )</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>( 40 )</td>
<td>12</td>
</tr>
<tr>
<td>( 10 )</td>
<td>( 10 )</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>( 20 )</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>( 30 )</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>( 40 )</td>
<td>12</td>
</tr>
<tr>
<td>( \infty )</td>
<td>( 10 )</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>( 20 )</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>( 30 )</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>( 40 )</td>
<td>12</td>
</tr>
</tbody>
</table>
response rate of 80%, the relative bias is 23% for \( M = 5 \). The relative bias depends very little on \( M \), being somewhat larger for large \( M \) because the variance is smaller.

In this case, an approximately unbiased estimator of the variance of the domain mean is

\[
\hat{V}_{M,D,n} = U_{M,D,n} + \{1 + M^{-1} - 2(r + m_{D})^{-1}(r - r_{D})\} B_{M,D,n}
\]

(4.5)

where \( U_{M,D,n} \) and \( B_{M,D,n} \) are the quantities of equation (1.2) computed for the domain mean.

To extend the results to multiple cells, assume a cell mean model with \( G \) cells. It can be shown that

\[
E\{ B_{M,D,n}(\hat{\theta}_D)|A, A_R \} = n^{-2} \sum_{g=1}^{G} m_{Dg}(1 + r_{g}^{-1}m_{Dg}) \sigma_g^2
\]

and

\[
\text{bias}\{ \hat{V}_{M,D,n}(\hat{\theta}_D)|A, A_R \} = 2n^{-2} \sum_{g=1}^{G} m_{Dg}(1 - r_{g}^{-1}r_{Dg}) \sigma_g^2
\]

where \( r_g \) is the number of respondents in cell \( g \), \( m_{Dg} \) is the number of non-respondents in cell \( g \) and in the domain, and \( r_{Dg} \) is the number of respondents in cell \( g \) and in the domain. Thus, assuming that \( \sigma_g^2 \equiv \sigma^2 \),

\[
\hat{V}_{M,D,n} = U_{M,D,n} + \left\{1 + \frac{1}{M} - 2 \sum_{g=1}^{G} \frac{m_{Dg}r_{g}^{-1}(r_g - r_{Dg})}{\sum_{g=1}^{G} m_{Dg}r_{g}^{-1}(r_g + m_{Dg})}\right\} B_{M,D,n}
\]

(4.6)

is approximately unbiased for the variance of the MI domain estimator. Note that estimator (4.6) reduces to the usual MI variance estimator for the full population parameter.

The bias-adjusted estimators in equations (4.5) and (4.6) are appropriate for the approximate Bayesian bootstrap imputation and for the normal distribution imputation that were described in Rubin and Schenker (1986). An extension to a weighted sample is discussed in the next section.

5. Linear regression models

A regression model is a natural model to use for imputing missing \( Y \)-values when a \( p \)-dimensional vector \( x \) of explanatory variables is always observed. Consider the model

\[
Y_i = x_i' \beta + e_i,
\]

(5.1)

where the \( e_i \)s are normally and independently distributed random variables with mean 0 and variance \( \sigma^2 \), and \( x_i \) are treated as fixed. Schenker and Welsh (1988) studied MI for model (5.1) assuming normal errors and simple random sampling. Assume, without loss of generality, that the first \( r \) units respond. Let \( y_r = (Y_1, Y_2, \ldots, Y_r)' \) and \( X_r = (x_1, x_2, \ldots, x_r)' \). The Schenker and Welsh imputed value is

\[
\eta_{j,k} = x_j' \hat{\beta}_i \cdot (k) + e_{j,k}^*,
\]

(5.2)

where the expected value of the \( \beta_i^* \) is

\[
\hat{\beta}_r = (X_r'X_r)^{-1}X_r'y_r
\]

and the \( (\beta_i^*, e_{j,k}^*) \) are constructed so that conditions (2.3), (2.4) and (3.6) are satisfied.

We consider a sample with weights \( \alpha_i = w_i, i = 1, 2, \ldots, n \), that are the inverses of the selection probabilities, possibly with non-response or calibration adjustments. By linear regression
theory, the $\hat{w}_{ij}$ of equation (3.14) is

$$\hat{w}_{ij} = \sum_{i \in A_R} \gamma_{ji} w_i \quad (5.3)$$

where $\gamma_{ij} = x_i'(X_i'X_r)^{-1}x_j$ and

$$\text{bias}(\hat{V}_{M,n} | A, A_R) = \left( \sum_{j \in A_M} w_j^2 - \sum_{i \in A_R} \sum_{j \in A_M} w_i w_j \gamma_{ij} \right) \sigma^2 + o(n^{-1}). \quad (5.4)$$

If $w_j = a'x_j$ for some vector $a$, then $\hat{w}_{li} = w_i$ and $\hat{V}_{M,n}$ is approximately unbiased for the variance of the imputed estimator. Thus, the vector of the weights must be in the column space of $X$ for the MI estimator of the variance of an estimated population total to be unbiased.

When the weights are not included in the regression model, a bias-adjusted MI variance estimator can be defined as

$$\tilde{V}_{M,n} = U_{M,n} + (1 + M^{-1}) B_{M,n} - \text{bias}(\hat{V}_{M,n}) \quad (5.5)$$

where

$$\text{bias}(\hat{V}_{M,n}) = 2 \sum_{j \in A_M} w_j^2 - \sum_{i \in A_R} \sum_{j \in A_M} w_i w_j \gamma_{ij} B_{M,n} \quad \sum_{j \in A_M} w_j^2 + \sum_{i \in A_M} \sum_{j \in A_M} w_i w_j \gamma_{ij} B_{M,n}$$

and

$$E(B_{M,n} | A, A_R) = \left( \sum_{j \in A_M} w_j^2 + \sum_{i \in A_M} \sum_{j \in A_M} \gamma_{ij} w_i w_j \right) \sigma^2 + o(n^{-1}). \quad (5.6)$$

The proof for equation (5.6) can be found in Kim (2004), equation (C.3).

Even if the weights are in the model, variance estimates for domains can be biased. The estimator of a domain total for an unequal probability sample is of the form

$$\hat{\theta}_{D,n} = \sum_{i \in D} \alpha_i Y_i$$

where $\alpha_i = w_i z_i$ and $z_i$ is as defined for equation (4.1). The conditional bias in the MI estimated variance of the domain estimator is, by equation (3.14),

$$\text{bias}(\hat{V}_{M,D,n} | A, A_R) = 2 \left( \sum_{i \in A_M} w_i^2 z_i - \sum_{i \in A_R} \sum_{j \in A_M} \gamma_{ij} w_i w_j z_i z_j \right) \sigma^2 + o(n^{-1}), \quad (5.7)$$

which reduces to equation (4.4) under simple random sampling. A bias-adjusted MI variance for the domain estimator is

$$\tilde{V}_{M,n} = U_{M,n} + (1 + M^{-1}) B_{M,n} - \text{bias}(\hat{V}_{M,n}) \quad (5.8)$$

where

$$\text{bias}(\hat{V}_{M,n}) = 2 \sum_{j \in A_M} w_j^2 z_j - \sum_{i \in A_R} \sum_{j \in A_M} \gamma_{ij} w_i w_j z_i z_j B_{M,n} \quad \sum_{j \in A_M} w_j^2 z_j + \sum_{i \in A_M} \sum_{j \in A_M} \gamma_{ij} w_i w_j z_i z_j B_{M,n}$$

If all weights are equal, then the bias in equation (5.7) is non-negative. However, if the respondents have larger weights than the non-respondents, then the bias can be negative. For example, suppose that the weights of the respondents are all equal to $w_0$ and that the weights of the non-respondents are all equal to $c w_0$. Then, the bias term equation (5.7) reduces to
bias(\hat{V}_{M,D,n}|A,A_R) = 2m_Dcw_0^2(c - r^{-1}_D)\sigma^2 + o(n^{-1})

and the bias is negative if c < r^{-1}_D.

For the MI estimator of variance to be approximately unbiased, w_iz_i must be a linear function of x_i. Given the multitude of domain analyses that are commonly conducted with survey data, it would not be feasible to include the term w_iz_i for all domains in an imputation scheme. As M → ∞, the imputed estimator becomes

\[ \hat{\theta}_{\infty,D,n} = \sum_{i\in A_R} w_iz_iY_i + \sum_{i\in A_M} w_iz_iX_i^r\hat{\beta}_r. \] (5.9)

The second term on the right-hand side of this equality is the part that can produce an improvement in the efficiency relative to the mean of the domain respondents because \( \hat{\beta}_r \) can be a function of observations outside the domain. If variables for domains are included in the model, there is no gain in the efficiency of \( \hat{\theta}_{\infty,D,n} \) relative to the mean of the domain respondents. See also Breidt et al. (1996).

Whereas it is possible to include the estimation weights for some preplanned domain estimators in the regression model, it is impossible to do so for all domains for which estimates could be computed. In particular, many domain estimates will not have been identified at the time that the imputation was carried out. Therefore, MI is not generally recommended for public use data files. Alternative imputation and variance estimation methods are discussed in Rao and Shao (1992), Särndal (1992), Brick et al. (2004), Kim and Fuller (2004) and Fuller and Kim (2005).

Acknowledgements

The authors acknowledge the National Center for Education Statistics, Institute for Education Sciences, for funding this research, and in particular the support of the Project Officer, Marilyn Seastrom. We also thank the Associate Editor and referees for their constructive comments.

Appendix A

A.1. Proof of lemma 1

By the definition of \( B_{M,n} \) and by condition (C.5),

\[ E(B_{M,n}|A,A_R) = E\{0.5M^{-1} (M-1)^{-1}\sum_{k=1}^M \sum_{l=1}^M (\hat{\theta}_{H(k)} - \hat{\theta}_{H(l)})^2 | A,A_R \} \]

\[ = 0.5 \text{var}(\hat{\theta}_{H(1)} - \hat{\theta}_{H(2)} | A,A_R). \] (A.1)

Also, by condition (C.5) again,

\[ \text{var}(\hat{\theta}_{M,n}|A,A_R) = (1-M^{-1}) \text{cov}(\hat{\theta}_{H(1)}, \hat{\theta}_{H(2)}| A,A_R) + M^{-1} \text{var}(\hat{\theta}_{H(1)}| A,A_R) \] (A.2)

and

\[ \text{cov}(\hat{\theta}_{M,n}, \hat{\theta}_n| A,A_R) = \text{cov}(\hat{\theta}_{H(1)}, \hat{\theta}_{H(2)} | A,A_R). \] (A.3)

Letting \( M \rightarrow \infty \) for both sides of equation (A.2),

\[ \text{var}(\hat{\theta}_{\infty,n}|A,A_R) = \text{cov}(\hat{\theta}_{H(1)}, \hat{\theta}_{H(2)} | A,A_R). \] (A.4)

By equations (A.3) and (A.4), the variance due to missingness is

\[ \text{var}(\hat{\theta}_{\infty,n} - \hat{\theta}_n| A,A_R) = \text{cov}(\hat{\theta}_{H(1)}, \hat{\theta}_{H(2)} | A,A_R) - 2 \text{cov}(\hat{\theta}_{H(1)}, \hat{\theta}_n | A,A_R) + \text{var}(\hat{\theta}_n | A,A_R) \]

\[ = \text{var}(\hat{\theta}_{H(1)} - \hat{\theta}_n | A,A_R) - 0.5 \text{var}(\hat{\theta}_{H(1)} - \hat{\theta}_{H(2)} | A,A_R). \] (A.5)
Therefore, by equations (A.1) and (A.5),
\[
E(B_{M,n}|A, A_R) - \text{var}(\hat{\theta}_{\infty,n} - \hat{\theta}_n|A, A_R) = E(\hat{\theta}_{R(1)} - \hat{\theta}_{R(2)}|A, A_R) - \text{var}(\hat{\theta}_{R(1)} - \hat{\theta}_n|A, A_R). \tag{A.6}
\]
By the definition of \(\hat{\theta}_{R(1)}\),
\[
\text{var}(\hat{\theta}_{R(1)} - \hat{\theta}_{R(2)}|A, A_R) = \text{var}\left\{ \sum_{i\in AM} \alpha_i(\eta_{i(1)} - \eta_{i(2)})|A, A_R \right\} \tag{A.7}
\]
and
\[
\text{var}(\hat{\theta}_{R(1)} - \hat{\theta}_n|A, A_R) = \text{var}\left( \sum_{i\in AM} \alpha_i\eta_{i(1)}|A, A_R \right) + \text{var}\left( \sum_{i\in AM} \alpha_i Y_i|A, A_R \right), \tag{A.8}
\]
where the equality holds because \(\eta_{i(1)}\) is a function of respondent values only. Therefore, equation (3.5) follows by inserting equations (A.7) and (A.8) into equation (A.6).

To show equation (3.7), note that the bias of \(B_{M,n}\) can be expressed as
\[
E(B_{M,n}|A, A_R) - \text{var}(\hat{\theta}_{\infty,n} - \hat{\theta}_n|A, A_R) = \sum_{i\in AM} \alpha_i^2 \{ \text{var}(\eta_{i(1)}|A, A_R) - \text{var}(Y_i|A, A_R) - 2 \text{cov}(\eta_{i(1)}, \eta_{i(2)}|A, A_R) \} + \sum_{i\in AM} \alpha_i \sum_{j\in AM, j\neq i} \alpha_j \{ \text{cov}(\eta_{i(1)}, \eta_{j(1)}|A, A_R) - \text{cov}(Y_i, Y_j|A, A_R) - 2 \text{cov}(\eta_{i(1)}, \eta_{j(2)}|A, A_R) \}. \tag{A.9}
\]
Under the superpopulation model with \(\text{cov}(Y_i, Y_j) = 0\) for \(i \neq j\), condition (C.4) implies that
\[
\text{var}(\eta_{i(1)}|A, A_R) = \text{var}(Y_i) + o(1) \tag{A.10}
\]
and
\[
\text{cov}(\eta_{i(1)}, \eta_{i(2)}|A, A_R) = o(1). \tag{A.11}
\]
Thus, by equations (A.10) and (A.11), the first summation in equation (A.9) is of smaller order than \(n^{-1}\). By equation (3.6), the second summation in equation (A.9) is of smaller order than \(n^{-1}\).

### A.2. Proof of lemma 2

By condition (C.5), for any \(M \geq 1\),
\[
\text{cov}(\hat{\theta}_{M,n}, \hat{\theta}_{\infty,n}|A, A_R) = \text{cov}(\hat{\theta}_{R(1)}, \hat{\theta}_{\infty,n}|A, A_R) = \text{cov}(\hat{\theta}_{R(1)}, \hat{\theta}_{R(2)}|A, A_R). \tag{A.12}
\]
By equations (A.2), (A.4) and (A.12),
\[
\text{var}(\hat{\theta}_{M,n} - \hat{\theta}_{\infty,n}|A, A_R) = M^{-1} \{ \text{var}(\hat{\theta}_{R(1)}|A, A_R) - \text{cov}(\hat{\theta}_{R(1)}, \hat{\theta}_{R(2)}|A, A_R) \}
\]
and the right-hand side of the equality is equal to \(E(M^{-1}B_{M,n}|A, A_R)\) by equation (A.1). Equality (3.9) directly follows from equations (A.3) and (A.12).

### A.3. Proof of theorem 1

Write the variance of the MI point estimator as
\[
\text{var}(\hat{\theta}_{M,n}) = \text{var}(\hat{\theta}_n) + \text{var}(\hat{\theta}_{M,n} - \hat{\theta}_n) + 2 \text{cov}(\hat{\theta}_n, \hat{\theta}_{M,n} - \hat{\theta}_n). \tag{A.13}
\]
By equations (2.2), (3.2) and (3.3),
\[
E(U_{M,n}) = \text{var}(\hat{\theta}_n) + o(n^{-1}). \tag{A.14}
\]
Also, by equations (2.4), (3.10), (3.8) and (3.9),

\[
(1 + M^{-1}) E(\hat{B}_{M,n}|A, A_R) = \text{var}(\hat{\theta}_{\infty,n} - \hat{\theta}_n|A, A_R) + \text{var}(\hat{\theta}_{M,n} - \hat{\theta}_{\infty,n}|A, A_R) + o(n^{-1})
\]

\[
= \text{var}(\hat{\theta}_{M,n} - \hat{\theta}_n|A, A_R) + o(n^{-1}), \tag{A.15}
\]

and, by condition (C.3), the conditional variance terms are all equal to the unconditional variances. It follows from equations (A.14) and (A.15) that \( \hat{V}_{M,n} \) estimates the sum of the first two terms on the right-hand side of equality (A.13) with a bias that is \( o(n^{-1}) \). Therefore the negative value of the third term on the right-hand side of equation (A.13) is the asymptotic bias. Because \( \hat{\theta}_{(k)} \) \( (k = 1, 2, \ldots, M) \) are identically distributed, the bias does not depend on \( M \).

References


