Biometrika (2016), **103**, 1, *pp.* 244–251
Printed in Great Britain

doi: 10.1093/biomet/asv073
Advance Access publication 5 February 2016

A note on multiple imputation for method of moments estimation

BY S. YANG

Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, Massachusetts 02115, U.S.A.

shuyang@hsph.harvard.edu

AND J. K. KIM

Department of Statistics, Iowa State University, Ames, Iowa 50011, U.S.A.

jkim@iastate.edu

SUMMARY

Multiple imputation is widely used for estimation in situations where there are missing data. Rubin (1987) provided an easily applicable formula for multiple imputation variance estimation, but its validity requires the congeniality condition of Meng (1994), which may not be satisfied for method of moments estimation. We give the asymptotic bias of Rubin's variance estimator when method of moments estimation is used in the complete-sample analysis for each imputed dataset. A new variance estimator based on over-imputation is proposed to provide asymptotically valid inference in this case.

Some key words: Bayesian method; Congeniality; Missingness at random; Proper imputation; Survey sampling.

1. INTRODUCTION

Imputation is often used to handle missing data. If imputed values are treated as if they were observed, variance estimates will generally be too small (Ford, 1983). To account for the uncertainty due to imputation, Rubin (1987, 1996) proposed multiple imputation.

Multiple imputation is set in a Bayesian framework. Rubin (1987) claimed that it can provide valid frequentist inference in various applications (e.g., Clogg et al., 1991). On the other hand, as discussed by Fay (1992), Kott (1995), Binder & Sun (1996), Fay (1996), Wang & Robins (1998), Robins & Wang (2000), Nielsen (2003) and Kim et al. (2006), the multiple imputation variance estimator can be inconsistent.

For multiple imputation inference to be valid, imputations must be proper (Rubin, 1987). A sufficient condition for this is the congeniality condition of Meng (1994), imposed on both the imputation model and the subsequent complete-sample analysis, which is quite restrictive for general-purpose estimation. Rubin's variance estimator is otherwise inconsistent. Kim (2011) pointed out that multiple imputation that is congenial for mean estimation is not necessarily congenial for proportion estimation, so some common statistical procedures, such as moment estimators, can be incompatible with multiple imputation.

In this paper, we characterize the asymptotic bias of Rubin's variance estimator when a method of moments estimator is used in the complete-sample analysis and discuss an alternative variance estimator that can provide asymptotically valid inference. The new variance estimator is compared with Rubin's variance estimator via simulations.

2. BASIC SET-UP

Suppose that the complete data consist of n observations $(x_1, y_1), \ldots, (x_n, y_n)$, which are independent realizations of a random vector (X, Y). For simplicity of presentation, assume that Y is a scalar response and X is a p-dimensional covariate. Suppose that x_i is fully observed and y_i is unobserved for some units in the sample. Without loss of generality, assume that the first r units of y_i are observed and the remaining $n-r$ units are missing. Let δ_i be the response indicator of y_i, i.e., $\delta_i = 1$ if y_i is observed and $\delta_i = 0$ otherwise. Let $y_{\text{obs}} = (y_1, \ldots, y_r)^{\mathrm{T}}$ and $X_n = (x_1, \ldots, x_n)$, and suppose that y_i is missing at random in the sense of Rubin (1976). The parameter of interest is $\eta = E\{g(Y)\}$, where $g(\cdot)$ is known. For example, if $g(y) = y$, then $\eta = E(Y)$ is the population mean of Y; and if $g(y) = I(y < 1)$, then $\eta = \text{pr}(Y < 1)$ is the population proportion of Y values less than 1.

Assume that the conditional density $f(y \mid x)$ belongs to a parametric class of models $f(y \mid x; \theta)$ $(\theta \in \Omega)$ but the marginal distribution of x is completely unspecified. To generate imputed values for missing outcomes from $f(y \mid x; \theta)$, we need to estimate the unknown parameter θ. The multiple imputation method adopts a Bayesian approach and proceeds in three steps.

Step 1. Create M complete datasets by filling in missing values with imputed values generated from the posterior predictive distribution. Specifically, to create the jth imputed dataset, first generate $\theta^{*(j)}$ from the posterior distribution $p(\theta \mid X_n, y_{\text{obs}})$, and then generate $y_i^{*(j)}$ from the imputation model $f(y \mid x_i; \theta^{*(j)})$ for each missing y_i.

Step 2. Apply the complete-sample estimation procedure to each imputed dataset. Let $\hat{\eta}^{(j)}$ be the complete-sample estimator of $\eta = E\{g(Y)\}$ applied to the jth imputed dataset and $\hat{V}^{(j)}$ the complete-sample variance estimator for $\hat{\eta}^{(j)}$.

Step 3. Use the combining rule of Rubin (1987) to summarize the results from the multiply imputed datasets. The multiple imputation estimator of η is $\hat{\eta}_{\text{MI}} = M^{-1}\sum_{j=1}^{M}\hat{\eta}^{(j)}$, and Rubin's variance estimator is

$$\hat{V}_{\text{MI}}(\hat{\eta}_{\text{MI}}) = W_M + (1 + M^{-1})B_M, \tag{1}$$

where $W_M = M^{-1}\sum_{j=1}^{M}\hat{V}^{(j)}$ and $B_M = (M-1)^{-1}\sum_{j=1}^{M}(\hat{\eta}^{(j)} - \hat{\eta}_{\text{MI}})^2$.

If the method of moments estimator of $\eta = E\{g(Y)\}$ is used in Step 2, the multiple imputation estimator of η becomes

$$\hat{\eta}_{\text{MI}} = M^{-1}\sum_{j=1}^{M}\hat{\eta}^{(j)} = n^{-1}\left\{\sum_{i=1}^{r} g(y_i) + \sum_{i=r+1}^{n} M^{-1}\sum_{j=1}^{M} g(y_i^{*(j)})\right\}, \tag{2}$$

where $\hat{\eta}^{(j)} = n^{-1}\{\sum_{i=1}^{r} g(y_i) + \sum_{i=r+1}^{n} g(y_i^{*(j)})\}$. To derive the frequentist properties of $\hat{\eta}_{\text{MI}}$, we rely on the Bernstein–von Mises theorem (van der Vaart, 2000, ch. 10), which states that under regularity conditions, the posterior distribution $p(\theta \mid X_n, y_{\text{obs}})$ is asymptotically normal with mean $\hat{\theta}$ and variance I_{obs}^{-1} almost surely, where $\hat{\theta}$ is the maximum likelihood estimator of θ and I_{obs}^{-1} is the inverse of the observed Fisher information matrix, $I_{\text{obs}} = -\sum_{i=1}^{r} \partial^2 \log f(y_i \mid x_i; \hat{\theta})/(\partial\theta\partial\theta^{\mathrm{T}})$. If $E\{g(Y) \mid x_i; \theta\}$ is sufficiently smooth in θ, conditional on the observed data, we have $\text{plim}_{M\to\infty} M^{-1}\sum_{j=1}^{M} g(y_i^{*(j)}) = E[E\{g(Y) \mid x_i; \theta^*\} \mid X_n, y_{\text{obs}}] \cong E\{g(Y) \mid x_i; \hat{\theta}\}$, where $A_n \cong B_n$ means $A_n = B_n + o_{\text{p}}(1)$ if A_n and B_n are random variables, or $A_n = B_n + o(1)$ if A_n and B_n are nonrandom variables. Therefore, as $M \to \infty$, $\hat{\eta}_{\text{MI}}$ converges to $\hat{\eta}_{\text{MI},\infty} = n^{-1}\{\sum_{i=1}^{r} y_i + \sum_{i=r+1}^{n} m(x_i; \hat{\theta})\}$, where $m(x; \theta) = E\{g(Y) \mid x; \theta\}$. The variance estimation of $\hat{\eta}_{\text{MI},\infty}$ needs to appropriately account for the uncertainty associated with estimating θ, and this is usually done by using linearization methods if the imputation models are known (Robins & Wang, 2000; Kim & Rao, 2009). In the multiple imputation procedure, the uncertainty in the estimation of θ is reflected in the variability between the multiply imputed datasets without referring to the imputation models. However, Rubin's variance estimator (1) requires restrictive conditions for valid inference.

3. Main result

Rubin's variance estimator is based on the decomposition

$$\text{var}(\hat{\eta}_{\text{MI}}) = \text{var}(\hat{\eta}_n) + \text{var}(\hat{\eta}_{\text{MI}} - \hat{\eta}_n) + 2\,\text{cov}(\hat{\eta}_{\text{MI}} - \hat{\eta}_n,\ \hat{\eta}_n), \tag{3}$$

where $\hat{\eta}_n$ is the complete-sample estimator of η. In Rubin's variance estimator (1), W_M estimates the first term of (3) and $(1+M^{-1})B_M$ estimates the second term of (3). Kim et al. (2006) proved that $E\{(1+M^{-1})B_M\} \cong \mathrm{var}(\hat{\eta}_{\mathrm{MI}} - \hat{\eta}_n)$ for a fairly general class of estimators. Hence, if the complete-sample variance estimator satisfies the condition $E(\hat{V}^{(j)}) \cong \mathrm{var}(\hat{\eta}_n)$ for $j=1,\ldots,M$, the bias of Rubin's variance estimator is

$$\mathrm{bias}(\hat{V}_{\mathrm{MI}}) \cong -2\,\mathrm{cov}(\hat{\eta}_{\mathrm{MI}} - \hat{\eta}_n, \hat{\eta}_n). \tag{4}$$

Thus, Rubin's variance estimator is asymptotically unbiased if $\mathrm{cov}(\hat{\eta}_{\mathrm{MI}} - \hat{\eta}_n,\ \hat{\eta}_n) \cong 0$; this was termed the congeniality condition by Meng (1994). However, this condition does not hold for some common estimators, such as method of moments estimators. Theorem 1 gives the asymptotic bias of Rubin's variance estimator as $M \to \infty$. Its proof is outlined in the Supplementary Material.

Theorem 1. *Let $\hat{\eta}_n = n^{-1}\sum_{i=1}^n g(y_i)$ be the method of moments estimator of $\eta = E\{g(Y)\}$ under complete response. Assume that $E(\hat{V}^{(j)}) \cong \mathrm{var}(\hat{\eta}_n)$ $(j=1,\ldots,M)$. Then, as $M\to\infty$, the bias of Rubin's variance estimator is*

$$\mathrm{bias}(\hat{V}_{\mathrm{MI}}) \cong 2n^{-1}(1-p)\big(E[\mathrm{var}\{g(Y)\mid X\}\mid \delta=0] - m_{\theta,0}^{\mathrm{T}}\mathcal{I}_\theta^{-1}m_{\theta,1}\big), \tag{5}$$

where $p=r/n$, $\mathcal{I}_\theta = -E\{\partial^2 \log f(Y\mid X;\theta)/(\partial\theta\partial\theta^{\mathrm{T}})\}$, $m(x;\theta)=E\{g(Y)\mid x;\theta\}$, $m_\theta(x) = \partial m(x;\theta)/\partial\theta$, $m_{\theta,0} = E\{m_\theta(X)\mid\delta=0\}$ and $m_{\theta,1} = E\{m_\theta(X)\mid\delta=1\}$.

Remark 1. Under missingness completely at random, (5) simplifies to

$$\mathrm{bias}(\hat{V}_{\mathrm{MI}}) \cong 2p(1-p)\{\mathrm{var}(\hat{\eta}_{r,\mathrm{MME}}) - \mathrm{var}(\hat{\eta}_{r,\mathrm{MLE}})\}, \tag{6}$$

where $\hat{\eta}_{r,\mathrm{MME}} = r^{-1}\sum_{i=1}^r g(y_i)$ and $\hat{\eta}_{r,\mathrm{MLE}} = r^{-1}\sum_{i=1}^r E\{g(Y)\mid x_i;\hat{\theta}\}$, because

$$\mathrm{var}(\hat{\eta}_{r,\mathrm{MME}}) = r^{-1}\mathrm{var}\{g(Y)\} = r^{-1}\mathrm{var}[E\{g(Y)\mid X\}] + r^{-1}E[\mathrm{var}\{g(Y)\mid X\}]$$

and

$$\mathrm{var}(\hat{\eta}_{r,\mathrm{MLE}}) \cong r^{-1}\mathrm{var}[E\{g(Y)\mid X\}] + r^{-1}m_\theta^{\mathrm{T}}\mathcal{I}_\theta^{-1}m_\theta$$

with $m_\theta = E\{m_\theta(X)\}$. Result (6) shows that Rubin's variance estimator is unbiased if and only if the method of moments estimator is as efficient as the maximum likelihood estimator, that is, $\mathrm{var}(\hat{\eta}_{r,\mathrm{MME}}) \cong \mathrm{var}(\hat{\eta}_{r,\mathrm{MLE}})$. Otherwise, Rubin's variance estimator is positively biased.

Remark 2. Under missingness at random, the bias of Rubin's variance estimator can be zero, positive or negative. Consider a simple linear regression model $Y = X^{\mathrm{T}}\beta + \epsilon$, where $\epsilon \sim N(0,\sigma^2)$. For $g(Y)=Y$, if X contains 1, then the method of moments estimator $n^{-1}\sum_{i=1}^n y_i$ is identical to the maximum likelihood estimator $n^{-1}\sum_{i=1}^n x_i^{\mathrm{T}}\hat{\beta}$, where $\hat{\beta}$ is the maximum likelihood estimator of β under complete response. Let $E_0(\cdot) = E(\cdot\mid\delta=0)$ and $E_1(\cdot)=E(\cdot\mid\delta=1)$. By (5) in Theorem 1 and direct calculation, considering that X contains 1, we have $\mathrm{bias}(\hat{V}_{\mathrm{MI}}) \cong 2n^{-1}(1-p)\sigma^2\{1 - E_0(X)^{\mathrm{T}}E_1(XX^{\mathrm{T}})^{-1}E_1(X)\} = 0$. This is consistent with Wang & Robins (1998) and Nielsen (2003). For a simple linear regression model with one covariate X and no intercept, the method of moments estimator is strictly less efficient than the maximum likelihood estimator (Matloff, 1981). The bias of Rubin's variance estimator is

$$\mathrm{bias}(\hat{V}_{\mathrm{MI}}) \cong 2n^{-1}(1-p)\sigma^2 E_1(X^2)^{-1}\{E_1(X^2) - E_0(X)E_1(X)\}, \tag{7}$$

which can be zero, positive or negative depending on the information for X in the respondent and nonrespondent groups.

4. Alternative variance estimator

We now discuss an alternative variance estimator that is unbiased regardless of whether the method of moments estimator or the maximum likelihood estimator is used as the complete-sample estimator in multiple imputation. We first decompose the multiple imputation estimator as $\hat{\eta}_{\rm MI} = \hat{\eta}_{\rm MI,\infty} + (\hat{\eta}_{\rm MI} - \hat{\eta}_{\rm MI,\infty})$. The two terms are uncorrelated because, conditional on the observed data, $\hat{\eta}_{\rm MI,\infty}$ is constant and the expectation of $\hat{\eta}_{\rm MI} - \hat{\eta}_{\rm MI,\infty}$ is zero. So

$$\mathrm{var}(\hat{\eta}_{\rm MI}) = \mathrm{var}(\hat{\eta}_{\rm MI,\infty}) + \mathrm{var}(\hat{\eta}_{\rm MI} - \hat{\eta}_{\rm MI,\infty}). \tag{8}$$

Note that $\mathrm{var}(\hat{\eta}_{\rm MI} - \hat{\eta}_{\rm MI,\infty})$ can be estimated by $M^{-1}B_M$ (Kim et al., 2006, Lemma 2). We now focus on estimating $\mathrm{var}(\hat{\eta}_{\rm MI,\infty})$ in (8); the details are given in the Supplementary Material. We show that

$$\mathrm{var}(\hat{\eta}_{\rm MI,\infty}) = V_1 + V_2 + V_3, \tag{9}$$

where $V_1 = n^{-1}\mathrm{var}\{m(X;\theta)\}$, $V_2 = r^{-1}m_\theta^{\rm T}\mathcal{I}_\theta^{-1}m_\theta$ and $V_3 = n^{-1}p\,\mathrm{var}[r^{-1}\sum_{i=1}^r\{g(y_i) - m(x_i;\hat{\theta})\}] \cong n^{-1}p(E[\mathrm{var}\{g(Y)\mid X\}\mid \delta=1] - m_{\theta,1}^{\rm T}\mathcal{I}_\theta^{-1}m_{\theta,1})$. The second term on the right-hand side of (9) reflects the variability associated with using the estimated rather than the true value of θ in the imputed values; the third term is the additional variance when the method of moments estimator is used instead of the maximum likelihood estimator.

For variance estimation, we use over-imputation: imputation is carried out for both the units with missing outcomes and those with observed outcomes. Over-imputation has been used to reduce disclosure risks (Reiter, 2004), as well as in model diagnostics for multiple imputation (Honaker et al., 2011; Blackwell et al., 2015).

Let $\bar{g}_i^* = M^{-1}\sum_{k=1}^M g(y_i^{*(k)})$ and $d_i^{(k)} = g(y_i^{*(k)}) - \bar{g}_i^*$ $(i=1,\ldots,n;\ k=1,\ldots,M)$. Define

$$C_{M,n} = \frac{1}{n^2(M-1)}\sum_{k=1}^M\sum_{i=1}^n (d_i^{*(k)})^2, \qquad C_{M,r} = \frac{1}{n^2(M-1)}\sum_{k=1}^M\sum_{i=1}^r (d_i^{*(k)})^2,$$
$$D_{M,n} = \frac{1}{M-1}\sum_{k=1}^M\left(\frac{1}{n}\sum_{i=1}^n d_i^{*(k)}\right)^2 - C_{M,n}, \quad D_{M,r} = \frac{1}{M-1}\sum_{k=1}^M\left(\frac{1}{n}\sum_{i=1}^r d_i^{*(k)}\right)^2 - C_{M,r}.$$

The key insight is based on the following observations:

$$E(C_{M,n}) \cong n^{-1}E[\mathrm{var}\{g(Y)\mid X\}], \qquad E(C_{M,r}) \cong n^{-1}pE[\mathrm{var}\{g(Y)\mid X\}\mid\delta=1],$$
$$E(D_{M,n}) \cong r^{-1}m_\theta^{\rm T}\mathcal{I}_\theta^{-1}m_\theta, \qquad E(D_{M,r}) \cong r^{-1}p^2 m_{\theta,1}^{\rm T}\mathcal{I}_\theta^{-1}m_{\theta,1}.$$

Therefore, the first term on the right-hand side of (9) can be estimated by

$$\hat{V}_1 = \frac{1}{n(n-1)}\left\{\sum_{i=1}^n \bar{g}_i^{*2} - \frac{1}{n}\left(\sum_{i=1}^n \bar{g}_i^*\right)^2\right\} - \frac{1}{M}C_{M,n}, \tag{10}$$

the second term can be estimated by

$$\hat{V}_2 = D_{M,n}, \tag{11}$$

and the third term can be estimated by

$$\hat{V}_3 \equiv C_{M,r} - D_{M,r}. \tag{12}$$

Combining the estimators of the three terms on the right-hand side of (9), we obtain the new multiple imputation variance estimator.

THEOREM 2. *Under the assumptions of Theorem 1, a multiple imputation variance estimator is*

$$\hat{V}_{\rm MI} = \hat{V}_1 + \hat{V}_2 + \hat{V}_3 + \frac{1}{M}B_M, \tag{13}$$

where $\hat{V}_1$, $\hat{V}_2$ and $\hat{V}_3$ are defined in (10)–(12) and B_M is the usual between-imputation variance in (1). As $n\to\infty$, the multiple imputation variance estimator is asymptotically unbiased for estimating the variance of the multiple imputation estimator in (2).

Remark 3. For small M, negative estimates of the variance may occur. In this case, we define $\hat{V}_1^+ = \max(\hat{V}_1, 0)$, $\hat{V}_2^+ = \max(\hat{V}_2, 0)$ and $\hat{V}_3^+ = \max(\hat{V}_3, 0)$, so that each estimate of nonnegative quantities is guaranteed to be nonnegative. The price we pay for setting negative values to zero is a slight increase in bias, but this disappears for large M. Notably, the upward bias is essentially negligible in our simulation study in § 5.

Remark 4. To account for the uncertainty in the variance estimator with a small to moderate imputation size, a $100(1-\alpha)\%$ interval estimate for η is $\hat{\eta}_{\mathrm{MI}} \pm t_{\nu,1-\alpha/2}\hat{V}_{\mathrm{MI}}^{1/2}$, where

$$\nu = \frac{(M-1)\hat{V}_{\mathrm{MI}}^2}{D_{M,n}^2 + D_{M,r}^2 + M^{-2}B_M^2 - 2p^2 D_{M,n}D_{M,r} + 2M^{-1}(1-p)^2 D_{M,n}B_M} \tag{14}$$

is an approximate number of degrees of freedom based on the Satterthwaite (1946) method; see the Supplementary Material.

Remark 5. The variance estimator (13) is also asymptotically unbiased when $\hat{\eta}_n$ is the maximum likelihood estimator of $\eta = E\{g(Y)\}$, so the proposed variance estimator is applicable whether the maximum likelihood or method of moments estimator is used. The price we pay for the better performance of our variance estimator is a slight increase in computational complexity and data storage space: it requires $M+1$ datasets, with M of them including the over-imputations and the last containing the original observed data. However, compared to the most common method of multiple imputation, it requires only one additional copy of imputations. When one's main concern is with valid inference, our proposed variance estimator based on over-imputation is preferred to Rubin's estimator. In addition, given over-imputations, the subsequent inference does not require knowledge of the imputation models. This is important because data analysts typically do not have access to all the information used for imputation. Our study would advocate over-imputation at the time of imputation, which not only allows imputers to assess the adequacy of the imputation models but also enables analysts to carry out valid inference without knowledge of the imputation models.

5. SIMULATION STUDY

We conducted a simulation study with two settings. For the first setting, 5000 Monte Carlo samples of size $n = 2000$ were independently generated from $Y_i = \beta X_i + e_i$, where $\beta = 0{\cdot}1$, $X_i \sim \exp(1)$ and $e_i \sim N(0, \sigma_e^2)$ with $\sigma_e^2 = 0{\cdot}5$. In the sample, we assume that X_i is fully observed but Y_i is not. Let δ_i be the response indicator of y_i and let $\delta_i \sim \mathrm{Ber}(p_i)$, where $p_i = 1/\{1+\exp(-\phi_0 - \phi_1 x_i)\}$. We consider two scenarios: (i) $(\phi_0, \phi_1) = (-1{\cdot}5, 2)$; (ii) $(\phi_0, \phi_1) = (3, -3)$, with the average response rate being about $0{\cdot}6$. The parameters of interest are $\eta_1 = E(Y)$ and $\eta_2 = \mathrm{pr}(Y < 0{\cdot}15)$. For multiple imputation, $M = 500$ imputed values were independently generated from the linear regression model using the Bayesian regression imputation procedure discussed in Schenker & Welsh (1998), where β and σ_e^2 are treated as independent with prior density proportional to σ_e^{-2}. For each imputed dataset, we employed the following complete-sample point estimators and variance estimators: $\hat{\eta}_{1,n} = n^{-1}\sum_{i=1}^n y_i$, $\hat{\eta}_{2,n} = n^{-1}\sum_{i=1}^n I(y_i < 0{\cdot}15)$, $\hat{V}(\hat{\eta}_{1,n}) = n^{-1}(n-1)^{-1}\sum_{i=1}^n (y_i - \hat{\eta}_{1,n})^2$ and $\hat{V}(\hat{\eta}_{2,n}) = (n-1)^{-1}\hat{\eta}_{2,n}(1-\hat{\eta}_{2,n})$. The relative bias of the variance estimator was calculated as $\{E(\hat{V}_{\mathrm{MI}}) - \mathrm{var}(\hat{\eta}_{\mathrm{MI}})\}/\mathrm{var}(\hat{\eta}_{\mathrm{MI}}) \times 100\%$. The $100(1-\alpha)\%$ confidence intervals were calculated as $(\hat{\eta}_{\mathrm{MI}} - t_{\nu,1-\alpha/2}\surd\hat{V}_{\mathrm{MI}},\ \hat{\eta}_{\mathrm{MI}} + t_{\nu,1-\alpha/2}\surd\hat{V}_{\mathrm{MI}})$, where $t_{\nu,1-\alpha/2}$ is the $100(1-\alpha/2)\%$ quantile of the t distribution with ν degrees of freedom. For Rubin's method, $\nu = (M-1)\lambda^{-2}$ with $\lambda = (1+M^{-1})B_M/\{W_M + (1+M^{-1})B_M\}$ (Barnard & Rubin, 1999). In our method, ν is as given in (14). The coverage was calculated as the percentage of Monte Carlo samples for which the true value falls within the confidence interval.

From Table 1 we can see that for $\eta_1 = E(Y)$, under scenario (i), the relative bias of Rubin's variance estimator is $94{\cdot}1\%$, consistent with (7) with $E_1(X^2) - E_0(X)^{\mathrm{T}}E_1(X) > 0$, where $E_1(X^2) = 3{\cdot}38$, $E_1(X) = 1{\cdot}45$ and $E_0(X) = 0{\cdot}48$. Under scenario (ii), the relative bias of Rubin's variance estimator is $-20{\cdot}1\%$, consistent with (7) with $E_1(X^2) - E_0(X)^{\mathrm{T}}E_1(X) < 0$, where $E_1(X^2) = 0{\cdot}37$, $E_1(X) = 0{\cdot}47$ and

Table 1. *Relative biases of two variance estimators together with the mean width and coverages of two interval estimates under two scenarios for the first simulation setting*

Scenario		Relative bias (%)		Mean width ($\times 10^3$) for 90% CI		Mean width ($\times 10^3$) for 95% CI		Coverage ($\times 10^2$) for 90% CI		Coverage ($\times 10^2$) for 95% CI	
		Rubin	B-C	Rubin	B-C	Rubin	B-C	Rubin	B-C	Rubin	B-C
(i)	η_1	94·1	−0·1	32	23	38	27	98	90	99	95
	η_2	114·6	−1·2	22	15	27	18	98	90	100	95
(ii)	η_1	−20·1	0·3	51	58	61	69	85	90	92	95
	η_2	−9·1	0·1	31	33	37	39	87	90	93	95

CI, confidence interval; η_1, $E(Y)$; η_2, $\text{pr}(Y < 0{\cdot}15)$; Rubin, the variance estimator of Rubin (1987); B-C, the proposed bias-corrected variance estimator.

Table 2. *Relative biases of two variance estimators together with the mean width and coverages of two interval estimates under two scenarios of missingness for the second simulation setting*

	M	Relative bias (%)		Mean width ($\times 10^3$) for 90% CI		Mean width ($\times 10^3$) for 95% CI		Coverage ($\times 10^2$) for 90% CI		Coverage ($\times 10^2$) for 95% CI	
		Rubin	B-C	Rubin	B-C	Rubin	B-C	Rubin	B-C	Rubin	B-C
					Missingness completely at random						
η_1	10	2·6	2·7	60	61	72	73	90	90	95	95
	30	1·9	2·5	60	61	72	72	90	90	95	95
η_2	10	29·7	2·1	22	19	26	23	94	90	98	95
	30	29·2	2·6	21	19	26	23	94	90	98	95
					Missingness at random						
η_1	10	1·9	2·1	59	60	71	71	90	90	95	95
	30	1·4	1·2	59	59	70	71	90	90	95	95
η_2	10	26·6	2·6	21	19	25	23	94	90	97	95
	30	26·2	1·9	21	19	25	23	94	90	97	95

CI, confidence interval; η_1, $E(Y)$; η_2, $\text{pr}(Y < 1)$; Rubin, the variance estimator of Rubin (1987); B-C, the proposed bias-corrected variance estimator.

$E_0(X) = 1{\cdot}73$. The empirical coverage of Rubin's method can be above or below the nominal coverage, but our variance estimator is essentially unbiased.

For the second simulation setting, 5000 Monte Carlo samples of size $n = 2000$ were independently generated from $Y_i = \beta_0 + \beta_1 X_i + e_i$, where $\beta = (\beta_0, \beta_1) = (3, -1)$, $X_i \sim N(2, 1)$ and $e_i \sim N(0, \sigma_e^2)$ with $\sigma_e^2 = 1$. The parameters of interest are $\eta_1 = E(Y)$ and $\eta_2 = \text{pr}(Y < 1)$. We considered two different factors. One is the response mechanism: missingness completely at random or missingness at random. For missingness completely at random, $\delta_i \sim \text{Ber}(0{\cdot}6)$; for missingness at random, $\delta_i \sim \text{Ber}(p_i)$ where $p_i = 1/\{1 + \exp(-\phi_0 - \phi_1 x_i)\}$ with $(\phi_0, \phi_1) = (0{\cdot}28, 0{\cdot}1)$ and with an average response rate of about 0·6. The other factor is the size of multiple imputation, $M = 10$ or 30.

Table 2 shows that Rubin's variance estimator is unbiased for $\eta_1 = E(Y)$, with absolute relative bias less than 2·6%, and our variance estimator is comparable, with absolute relative bias less than 3·0%. Rubin's variance estimator is biased upward for $\eta_2 = \text{pr}(Y < 1)$, with absolute relative bias as high as 29·7%, whereas our variance estimator reduces the absolute relative bias to less than 2·7%. As regards confidence interval estimates, for $\eta_1 = E(Y)$, the average length of confidence intervals from our method is slightly larger than that for Rubin's method, because (14) produces a smaller number of degrees of freedom. However, for $\eta_2 = \text{pr}(Y < 1)$, the average length of confidence intervals from our method is shorter even with a smaller number of degrees of freedom, due to the overestimation of Rubin's method. Rubin's method provides good empirical coverage for $\eta_1 = E(Y)$, but the empirical coverage for $\eta_2 = \text{pr}(Y < 1)$ reaches 94% for 90% confidence intervals and 98% for 95% confidence intervals, due to variance overestimation.

In contrast, our method provides more accurate coverage of confidence intervals for both $\eta_1 = E(Y)$ and $\eta_2 = \mathrm{pr}(Y < 1)$.

6. DISCUSSION

Our method can be extended to a more general class of parameters η obtained by solving estimating equations $\sum_{i=1}^{n} U(\eta; x_i, y_i) = 0$, such as the mean of y, the proportion of y less than q, the pth quantile, regression coefficients, and domain means. A similar approach could be used to characterize the bias of Rubin's variance estimator and to develop a bias-corrected variance estimator.

Another extension would be to develop unbiased variance estimation for the vector case of η with $q > 1$ components. As in the scalar case, we can construct multivariate analogues of the multiple imputation estimator and the variance estimator; however, finding an adequate reference distribution for the statistic $(\hat{\eta}_{\mathrm{MI}} - \eta)^{\mathrm{T}} \hat{V}_{\mathrm{MI}}^{-1} (\hat{\eta}_{\mathrm{MI}} - \eta)/q$ is harder than in the scalar case. One potential solution is to make a simplifying assumption that the fraction of missing information is equal for all the components of η, as discussed in Li et al. (1994) and Xie & Meng (2015).

ACKNOWLEDGEMENT

We are grateful to Xianchao Xie, Xiao-Li Meng, the editors and four referees for their valuable comments, which have helped to improve this paper. The research of the second author was partially supported by the U.S. National Science Foundation and by a cooperative agreement between the U.S. Department of Agriculture Natural Resources Conservation Service and Iowa State University.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes the proofs of Theorems 1 and 2, verification of unbiasedness of the new variance estimator when $\hat{\eta}_n$ is the maximum likelihood estimator of $\eta = E\{g(Y)\}$, and derivation of the approximate number of degrees of freedom.

REFERENCES

BARNARD, J. & RUBIN, D. B. (1999). Small-sample degrees of freedom with multiple imputation. *Biometrika* **86**, 948–55.

BINDER, D. A. & SUN, W. (1996). Frequency valid multiple imputation for surveys with a complex design. In *Proc. Surv. Res. Meth. Sect. ASA*. American Statistical Association, pp. 281–6.

BLACKWELL, M., HONAKER, J. & KING, G. (2015). A unified approach to measurement error and missing data: Details and extensions. *Sociol. Meth. Res.* doi:10.1177/0049124115589052, 1–28.

CLOGG, C. C., RUBIN, D. B., SCHENKER, N., SCHULTZ, B. & WEIDMAN, L. (1991). Multiple imputation of industry and occupation codes in census public-use samples using Bayesian logistic regression. *J. Am. Statist. Assoc.* **86**, 68–78.

FAY, R. E. (1992). When are inferences from multiple imputation valid? In *Proc. Survey Res. Meth. Sect. ASA*. American Statistical Association, pp. 227–32.

FAY, R. E. (1996). Alternative paradigms for the analysis of imputed survey data. *J. Am. Statist. Assoc.* **91**, 490–8.

FORD, B. L. (1983). An overview of hot-deck procedures. In *Incomplete Data in Sample Surveys*, W. Madow, H. Niseelson & I. Olkin, eds., vol. 2. New York: Academic Press, pp. 185–207.

HONAKER, J., KING, G. & BLACKWELL, M. (2011). Amelia II: A program for missing data. *J. Statist. Comp. Software* **45**, 1–47.

KIM, J. K. (2011). Parametric fractional imputation for missing data analysis. *Biometrika* **98**, 119–32.

KIM, J. K. & RAO, J. N. K. (2009). A unified approach to linearization variance estimation from survey data after imputation for item nonresponse. *Biometrika* **96**, 917–32.

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.

KOTT, P. S. (1995). A paradox of multiple imputation. In *Proc. Survey Res. Meth. Sect. ASA*. American Statistical Association, pp. 384–9.

LI, K. H., RAGHUNATHAN, T. E. & RUBIN, D. B. (1994). Large-sample significance levels from multiply imputed data using moment-based statistics and an F reference distribution. *J. Am. Statist. Assoc.* **86**, 1065–73.

MATLOFF, N. S. (1981). Use of regression functions for improved estimation of means. *Biometrika* **68**, 685–9.

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

NIELSEN, S. F. (2003). Proper and improper multiple imputation. *Int. Statist. Rev.* **71**, 593–607.

REITER, J. P. (2004). Simultaneous use of multiple imputation for missing data and disclosure limitation. *Survey Methodol.* **30**, 235–42.

ROBINS, J. M. & WANG, N. (2000). Inference for imputation estimators. *Biometrika* **87**, 113–24.

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.

SATTERTHWAITE, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bull.* **2**, 110–4.

SCHENKER, N. & WELSH, A. H. (1998). Asymptotic results for multiple imputation. *Ann. Statist.* **16**, 1550–66.

VAN DER VAART, A. W. (2000). *Asymptotic Statistics*. Cambridge: Cambridge University Press.

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

XIE, X. & MENG, X. L. (2015). Dissecting multiple imputation from a multi-phase inference perspective: What happens when God's, imputer's and analyst's models are uncongenial? *Statist. Sinica* doi: 10.5705/ss.2014.067.

[*Received January* 2015. *Revised December* 2015]