2-2016

A note on multiple imputation for method of moments estimation

Shu Yang
Harvard University

Jae Kwang Kim
Iowa State University, jkim@iastate.edu

Biometrika (20xx), **xx**, x, *pp*. 1–8

Printed in Great Britain

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

J. K. KIM

Department of Statistics, Iowa State University, Ames, Iowa 50010, U.S.A.

jkim@iastate.edu

SUMMARY

Multiple imputation is a popular imputation method for general purpose estimation. Rubin (1987) provided an easily applicable formula for the variance estimation of multiple imputation. However, the validity of the multiple imputation inference requires the congeniality condition of Meng (1994), which is not necessarily satisfied for method of moments estimation. This paper presents the asymptotic bias of Rubin's variance estimator when the method of moments estimator is used as a complete-sample estimator in the multiple imputation procedure. A new variance estimator based on over-imputation is proposed to provide asymptotically valid inference for method of moments estimation.

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

1. INTRODUCTION

Imputation is often used to handle missing data. For inference, if imputed values are treated as if they were observed, variance estimates will generally be underestimates (Ford, 1983). To account for the uncertainty due to imputation, Rubin (1987, 1996) proposed multiple imputation which creates multiply completed datasets to allow assessment of imputation variability.

Multiple imputation is motivated in a Bayesian framework; however, its frequentist validity is controversial. Rubin (1987) claimed that multiple imputation can provide valid frequentist inference in various applications (for example, Clogg et al., 1991). On the other hand, as discussed by Fay (1992), Kott (1995), Fay (1996), Binder & Sun (1996), Wang & Robins (1998), Robins & Wang (2000), Nielsen (2003), and Kim et al. (2006), the multiple imputation variance estimator is not always consistent.

For multiple imputation inference to be valid, imputations must be proper (Rubin, 1987). A sufficient condition is given by Meng (1994), the so-called congeniality condition, imposed on both the imputation model and the form of subsequent complete-sample analyses, 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. Therefore, some common statistical procedures, such as the method of moments estimators, can be incompatible with the multiple imputation framework.

In this paper, we characterize the asymptotic bias of Rubin's variance estimator when the method of moments estimator is used in the complete-sample analysis. We also discuss an alternative variance estimator that can provide asymptotically valid inference for method of moments estimation. The new variance estimator is compared with Rubin's variance estimator through two limited simulation studies in §5.

2. BASIC SETUP

Suppose that the sample consists of n observations $(x_1, y_1), \ldots, (x_n, y_n)$, which is an independent realization of a random vector (X, Y). For simplicity of presentation, assume that Y is a scalar outcome variable and X is a p-dimensional covariate. Suppose that x_i is fully observed and y_i is not fully observed for all units in the sample. Without loss of generality, assume the first r units of y_i are observed and the remaining $n - r$ units of y_i are missing. Let δ_i be the response indicator of y_i, that is, $\delta_i = 1$ if y_i is observed and $\delta_i = 0$ otherwise. Denote $y_{\rm obs} = (y_1, \ldots, y_r)^T$ and $X_n = (x_1, \ldots, x_n)$. We further assume that the missing mechanism is missing at random in the sense of Rubin (1976). The parameter of interest is $\eta = E\{g(Y)\}$, where $g(\cdot)$ is a known function. For example, if $g(y) = y$, then $\eta = E(Y)$ is the population mean of Y, and if $g(y) = I(y < 1)$, then $\eta = {\rm pr}(Y < 1)$ is the population proportion of Y less than 1.

Assume that the conditional density $f(y \mid x)$ belongs to a parametric class of models indexed by θ such that $f(y \mid x) = f(y \mid x; \theta)$ for some $\theta \in \Omega$ and 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, either by likelihood-based methods or by Bayesian methods. The multiple imputation procedure employs a Bayesian approach to deal with the unknown parameter θ, which unfolds in three steps:

Step 1. (Imputation) 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 $p(\theta \mid X_n, y_{\rm obs})$, and then generate $y_i^{*(j)}$ from the imputation model $f(y \mid x_i; \theta^{*(j)})$ for each missing y_i.

Step 2. (Analysis) Apply the user's 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 $\hat{V}^{(j)}$ be the complete-sample variance estimator of $\hat{\eta}^{(j)}$.

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

$$\hat{V}_{\rm MI}(\hat{\eta}_{\rm MI}) = W_M + \left(1 + M^{-1}\right) B_M, \qquad (1)$$

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

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

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

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 property of $\hat{\eta}_{\rm MI}$, we rely on the Bernstein-von Mises theorem (van der Vaart, 2000; Chapter 10), which claims that under regularity conditions and conditional on the observed data, the posterior distribution $p(\theta \mid X_n, y_{\rm obs})$ converges to a normal distribution with mean $\hat\theta$ and variance $I_{\rm obs}^{-1}$, where $\hat\theta$ is the maximum likelihood estimator of θ from the observed data and $I_{\rm obs}^{-1}$ is the inverse of the observed Fisher information matrix with $I_{\rm obs} = -\sum_{i=1}^r \partial^2 \log f(y_i \mid x_i; \hat\theta)/\partial\theta\partial\theta^T$. As a result, assume that $E\{g(Y) \mid x_i; \theta\}$ is sufficiently smooth in θ, conditional on the observed data, we have $p\lim_{M\to\infty} M^{-1}\sum_{j=1}^M g(y_i^{*(j)}) = E\left[E\{g(Y) \mid x_i; \theta^*\} \mid X_n, y_{\rm obs}\right] \cong E\{g(Y) \mid x_i; \hat\theta\}$, where $A_n \cong B_n$ means $A_n = B_n + o_p(1)$. Therefore, for $M \to \infty$, $\hat\eta_{\rm MI}$ converges to $\hat\eta_{\rm MI,\infty} = n^{-1}\{\sum_{i=1}^r y_i + \sum_{i=r+1}^n m(x_i; \hat\theta)\}$, where $m(x;\theta) = E\{g(Y) \mid x; \theta\}$. The variance estimation of $\hat\eta_{\rm MI,\infty}$ needs to appropriately account for the uncertainty associated with the estimate of θ, which is usually done using linearization methods if the imputation models are known (Robins & Wang, 2000; Kim & Rao, 2009). In the multiple imputation procedure, this is characterized 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, which we discuss in the next section.

3. Main Result

Rubin's variance estimator is based on the following decomposition,

$$\mathrm{var}(\hat\eta_{\rm MI}) = \mathrm{var}(\hat\eta_n) + \mathrm{var}(\hat\eta_{\rm MI} - \hat\eta_n) + 2\mathrm{cov}(\hat\eta_{\rm MI} - \hat\eta_n, \hat\eta_n), \tag{3}$$

where $\hat\eta_n$ is the complete-sample estimator of η. Basically, 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). In particular, Kim et al. (2006) proved that $E\{(1 + M^{-1})B_M\} \cong \mathrm{var}(\hat\eta_{\rm MI} - \hat\eta_n)$ for a fairly general class of estimators. Thus, 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

$$\mathrm{bias}(\hat V_{\rm MI}) \cong -2\mathrm{cov}(\hat\eta_{\rm MI} - \hat\eta_n, \hat\eta_n). \tag{4}$$

Rubin's variance estimator is asymptotically unbiased if $\mathrm{cov}(\hat\eta_{\rm MI} - \hat\eta_n, \hat\eta_n) \cong 0$, which is called the congeniality condition by Meng (1994). However, the congeniality condition does not hold for some common estimators such as the method of moments estimators. Theorem 1 gives this asymptotic bias of Rubin's variance estimator for $M \to \infty$, with the proof outlined in the online 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 $E(\hat V^{(j)}) \cong \mathrm{var}(\hat\eta_n)$ holds for $j = 1, \ldots, M$. Then for $M \to \infty$, the bias of Rubin's variance estimator is*

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

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

Remark 1. Under missing completely at random, the bias in (5) simplifies to

$$\mathrm{bias}(\hat V_{\rm MI}) \cong 2p(1-p)\{\mathrm{var}(\hat\eta_{r,\rm MME}) - \mathrm{var}(\hat\eta_{r,\rm MLE})\}, \tag{6}$$

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

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

and

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

where $\dot{m}_\theta = E\{\dot{m}_\theta(X)\}$. Result (6) explicitly 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 missing at random, the bias of Rubin's variance estimator can be zero, positive or negative. Consider a simple linear regression model $Y = X^T\beta + \epsilon$, 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 $n^{-1}\sum_{i=1}^n x_i^T\hat{\beta}$ with $\hat{\beta}$ being the maximum likelihood estimator of β under complete response. By Theorem 1, let $E_0(\cdot) = E(\cdot \mid \delta = 0)$ and $E_1(\cdot) = E(\cdot \mid \delta = 1)$, the bias of Rubin's variance estimator in (5) is $\mathrm{bias}(\hat{V}_{\mathrm{MI}}) \cong 2n^{-1}(1 - p)\sigma^2\{1 - E_0(X)^T E_1(XX^T)^{-1}E_1(X)\} = 0$, by direct calculation considering that X contains 1. This is consistent with the theory in Wang & Robins (1998) and Nielsen (2003). Now consider a simple linear regression model which contains one covariate X and no intercept, then the method of moments estimator is strictly less efficient than the maximum likelihood estimator (Matloff, 1981). The bias of Rubin's variance estimator is

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

which can be zero, positive or negative depending on the information of X in the respondent and non-respondent groups. See the first simulation study in §5.

4. ALTERNATIVE VARIANCE ESTIMATION

In this section, we consider an alternative variance estimation method that leads to an unbiased variance estimator for multiple imputation regardless of whether the method of moments estimator or the maximum likelihood estimator is used as the complete-sample estimator in the multiple imputation procedure. We first decompose the multiple imputation estimator as, $\hat{\eta}_{\mathrm{MI}} = \hat{\eta}_{\mathrm{MI},\infty} + (\hat{\eta}_{\mathrm{MI}} - \hat{\eta}_{\mathrm{MI},\infty})$. The two terms are uncorrelated using the law of total covariance and the fact that $\hat{\eta}_{\mathrm{MI},\infty}$ is the conditional expectation of $\hat{\eta}_{\mathrm{MI}}$, conditional on the observed data. Therefore, we have

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

Note that $\mathrm{var}(\hat{\eta}_{\mathrm{MI}} - \hat{\eta}_{\mathrm{MI},\infty})$ can be estimated by $M^{-1}B_M$ (Kim et al., 2006; Lemma 2). We now focus on estimating $\mathrm{var}(\hat{\eta}_{\mathrm{MI},\infty})$ in (8). For simplicity of presentation, all details of derivation are to be found in supplementary material. We show that the variance of $\hat{\eta}_{\mathrm{MI},\infty}$ is a sum of two terms,

$$\mathrm{var}\left(\hat{\eta}_{\mathrm{MI},\infty}\right) = n^{-1}V_1 + r^{-1}V_2, \quad (9)$$

where $V_1 = \mathrm{var}\{g(Y)\} - (1-p)E[\mathrm{var}\{g(Y) \mid X\} \mid \delta = 0]$, and $V_2 = \dot{m}_\theta^T\mathcal{I}_\theta^{-1}\dot{m}_\theta - p^2\dot{m}_{\theta,1}^T\mathcal{I}_\theta^{-1}\dot{m}_{\theta,1}$.

The first term, $n^{-1}V_1$, is the variance of the sample mean of $g(y_i) - (1-\delta_i)\{g(y_i) - m(x_i;\theta)\}$. To estimate this term, consider $W_M = M^{-1}\sum_{j=1}^M \hat{V}^{(j)}$ as in (1), and

$$C_M = \frac{1}{n^2(M-1)}\sum_{k=1}^M\sum_{i=r+1}^n\left\{g(y_i^{*(k)}) - \frac{1}{M}\sum_{k=1}^M g(y_i^{*(k)})\right\}^2. \quad (10)$$

We have $E\{W_M\} \cong n^{-1}\mathrm{var}\{g(Y)\}$ and $E(C_M) \cong n^{-1}(1-p)E[\mathrm{var}\{g(Y) \mid X\} \mid \delta = 0]$. Therefore, the first term $n^{-1}V_1$ can be estimated by $\tilde{W}_M = W_M - C_M$. By the strong law of large numbers, $\mathrm{pr}(\tilde{W}_M \geq 0) \to 1$ as $n \to \infty$.

The second term, $r^{-1}V_2$, reflects the variability associated with the estimated value of θ instead of the true value θ in the imputed values. To estimate this term, we use over-imputation in the sense that the imputation is carried out not only for the units with missing outcomes, but also for the units with observed outcomes. Over-imputation has been used in model diagnostics for multiple imputation (Honaker et al., 2010; Blackwell et al., 2015). Let $d_i^{(k)} = g(y_i^{*(k)}) - M^{-1}\sum_{l=1}^{M} g(y_i^{*(l)})$ for $i = 1, \ldots, n$ and $k = 1, \ldots, M$. Define $D_{M,n} = (M-1)^{-1}\sum_{k=1}^{M}(n^{-1}\sum_{i=1}^{n} d_i^{*(k)})^2 - (M-1)^{-1}\sum_{k=1}^{M} n^{-2}\sum_{i=1}^{n}(d_i^{*(k)})^2$, and $D_{M,r} = (M-1)^{-1}\sum_{k=1}^{M}(n^{-1}\sum_{i=1}^{r} d_i^{*(k)})^2 - (M-1)^{-1}\sum_{k=1}^{M} n^{-2}\sum_{i=1}^{r}(d_i^{*(k)})^2$. The key insight is based on the following observations: $E(D_{M,n}) \cong r^{-1}\dot{m}_\theta^T \mathcal{I}_\theta^{-1}\dot{m}_\theta$ and $E(D_{M,r}) \cong r^{-1}p^2\dot{m}_{\theta,1}^T\mathcal{I}_\theta^{-1}\dot{m}_{\theta,1}$; therefore, the second term of (9) can be estimated by $D_M = D_{M,n} - D_{M,r}$. Combining the estimators of the two terms in (9), we have the new multiple imputation variance estimator, given in the following theorem.

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

$$\hat{V}_{\mathrm{MI}} = \tilde{W}_M + D_M + M^{-1}B_M, \tag{11}$$

where $\tilde{W}_M = W_M - C_M$, with C_M defined in (10) and B_M being the usual between-imputation variance in (1). $\hat{V}_{\mathrm{MI}}$ is asymptotically unbiased for estimating the variance of the multiple imputation estimator in (2) as $n \to \infty$.

Remark 3. 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 $\hat{\eta}_{\mathrm{MI}} \pm t_{\mathrm{df},1-\alpha/2}\sqrt{\hat{V}_{\mathrm{MI}}}$, where df is an approximate number of degrees of freedom based on Satterthwaite's method (1946) given in supplementary material. From simulation studies, we find that using $df = M-1$ gives similar satisfactory results as using the formula we provided. As a practical matter, $df = M-1$ is preferred.

Remark 4. The proposed variance estimator in (11) is also asymptotically unbiased when $\hat{\eta}_n$ is the maximum likelihood estimator of $\eta = E\{g(Y)\}$ (see supplementary material for proof). Therefore, the proposed variance estimator is applicable regardless of whether the maximum likelihood estimator or the method of moments estimator is used for the complete-sample estimator. The price we pay for the better performance of our variance estimator is an increase in computational complexity and data storage space, which requires $M+1$ datasets, with M of them including the over-imputations and the last one containing the original observed data. However, when one's concern is with valid inference of multiple imputation, as in this paper, our proposed variance estimator based on over-imputation is preferred over that of Rubin's. In addition, given over-imputations, the subsequent inference does not require the knowledge of the imputation models. This is important because data analysts typically do not have access to all the information that the imputers used for imputation. Our study would promote the use of over-imputation at the time of imputation, which not only allows the imputers to assess the adequacy of the imputation models, but also enables the analysts to carry out valid inference without knowledge of the imputation models.

5. SIMULATION STUDY

To test our theory, we conduct two limited simulation studies. In the first simulation, $5,000$ Monte Carlo samples of size $n = 2,000$ are independently generated from $Y_i = \beta X_i + e_i$, where $\beta = 0.1$, $X_i \sim \exp(1)$ and $e_i \sim N(0, \sigma_e^2)$ with $\sigma_e^2 = 0.5$. 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 $\delta_i \sim \mathrm{Bernoulli}(p_i)$, where $p_i = 1/\{1 + \exp(-\phi_0 - \phi_1 x_i)\}$. We consider two scenarios: (i) $(\phi_0, \phi_1) = (-1.5, 2)$ and (ii) $(\phi_0, \phi_1) = (3, -3)$, with the average response rate about 0.6. The parameters of interest are $\eta_1 = E(Y)$ and $\eta_2 = \mathrm{pr}(Y < 0.15)$. For multiple imputation, $M = 500$ imputed values are 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}. In each imputed dataset, we adopt the following complete-sample point estimators and variance estimators: $\hat{\eta}_{1,n} = n^{-1}\sum_{i=1}^n y_i$, $\hat{\eta}_{2,n} = n^{-1}\sum_{i=1}^n I(y_i < 0.15)$, $\hat{V}(\hat{\eta}_{1,n}) = n^{-1}(n-1)^{-1}\sum_{i=1}^n (y_i - \hat{\eta}_{1,n})^2$, 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 is 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 are calculated as $(\hat{\eta}_{\mathrm{MI}} - t_{\nu,1-\alpha/2}\sqrt{\hat{V}_{\mathrm{MI}}}, \hat{\eta}_{\mathrm{MI}} + t_{\nu,1-\alpha/2}\sqrt{\hat{V}_{\mathrm{MI}}})$, 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, $\nu = \nu_1\nu_2/(\nu_1 + \nu_2)$ with $\nu_1 = (M-1)\lambda^{-2}$, $\nu_2 = (\nu_{\mathrm{com}} + 1)(\nu_{\mathrm{com}} + 3)^{-1}\nu_{\mathrm{com}}(1-\lambda)$, $\nu_{\mathrm{com}} = n - 3$, and $\lambda = (1 + M^{-1})B_M/\{W_M + (1 + M^{-1})B_M\}$ (Barnard & Rubin, 1999). In our new method, $\nu = M - 1$. The coverage is calculated as the percentage of Monte Carlo samples where the estimate falls within the confidence interval.

From Table 1, for $\eta_1 = E(Y)$, under scenario (i), the relative bias of Rubin's variance estimator is 96.8%, which is consistent with our result in (7) with $E_1(X^2) - E_0(X)^T E_1(X) > 0$, where $E_1(X^2) = 3.38$, $E_1(X) = 1.45$, and $E_0(X) = 0.48$. Under scenario (ii), the relative bias of Rubin's variance estimator is -19.8%, which is consistent with our result in (7) with $E_1(X^2) - E_0(X)^T E_1(X) < 0$, where $E_1(X^2) = 0.37$, $E_1(X) = 0.47$, and $E_0(X) = 1.73$. The empirical coverage for Rubin's method can be over or below the nominal coverage due to variance overestimation or underestimation. On the other hand, the new variance estimator is essentially unbiased for these scenarios.

In the second simulation, $5,000$ Monte Carlo samples of size $n = 200$ are 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 $\eta_2 = \mathrm{pr}(Y < 1)$. We consider two different factors for simulation. One is the response mechanism: missing completely at random and missing at random. For missing completely at random, $\delta_i \sim \mathrm{Bernoulli}(0.6)$. For missing at random, $\delta_i \sim \mathrm{Bernoulli}(p_i)$, where $p_i = 1/\{1 + \exp(-\phi_0 - \phi_1 x_i)\}$ and $(\phi_0, \phi_1) = (0.28, 0.1)$ with the average response rate about 0.6. The other factor is the size of multiple imputation, with two levels $M = 10$ and $M = 30$.

From Table 2, regarding the relative bias, Rubin's variance estimator is unbiased for $\eta_1 = E(Y)$, with absolute relative bias of less than 1%, and our new variance estimator is comparable with Rubin's variance estimator with absolute relative bias of less than 1.68%. Rubin's variance estimator is biased upward for $\eta_2 = \mathrm{pr}(Y < 1)$, with absolute relative bias as high as 24%; whereas our new variance estimator reduces absolute relative bias to less than 1.74%. Regarding confidence interval estimates, for $\eta_1 = E(Y)$, the confidence interval calculated from our new method is slightly wider than that from Rubin's method, because our new method uses a smaller number of degrees of freedom in the t distribution. However, for $\eta_2 = \mathrm{pr}(Y < 1)$, the confidence interval calculated from our new method is narrower than that from Rubin's method even with a smaller number of degrees of freedom in the t distribution, due to the overestimation in Rubin's

Table 1. *Relative biases of two variance estimators and mean width and coverages of two interval estimates under two scenarios in simulation one*

		Relative bias (%)		Mean Width for 90% C.I.		Mean Width for 95% C.I.		Coverage for 90% C.I.		Coverage for 95% C.I.	
Scenario		Rubin	New	Rubin	New	Rubin	New	Rubin	New	Rubin	New
1	η_1	96.8	0.7	0.032	0.023	0.038	0.027	0.98	0.90	0.99	0.95
	η_2	123.7	2.9	0.022	0.015	0.027	0.018	0.98	0.91	1.00	0.95
2	η_1	-19.8	0.4	0.051	0.058	0.061	0.069	0.85	0.90	0.91	0.95
	η_2	-9.6	-0.4	0.031	0.033	0.037	0.039	0.87	0.90	0.93	0.95

C.I., confidence interval; $\eta_1 = E(Y)$; $\eta_2 = \text{pr}(Y < 0.15)$; Rubin/New, Rubin's/New variance estimator.

Table 2. *Relative biases of two variance estimators and mean width and coverages of two interval estimates under two scenarios of missingness in simulation two*

		Relative Bias (%)		Mean Width for 90% C.I.		Mean Width for 95% C.I.		Coverage for 90% C.I.		Coverage for 95% C.I.	
	M	Rubin	New	Rubin	New	Rubin	New	Rubin	New	Rubin	New
					Missing completely at random						
η_1	10	-0.9	-1.58	0.20	0.211	0.24	0.25	0.90	0.90	0.95	0.95
	30	-0.6	-1.68	0.192	0.196	0.230	0.235	0.90	0.90	0.95	0.95
η_2	10	22.7	-1.14	0.069	0.067	0.083	0.083	0.95	0.90	0.98	0.95
	30	23.8	-1.23	0.068	0.062	0.082	0.075	0.94	0.90	0.98	0.95
					Missing at random						
η_1	10	-1.0	-1.48	0.19	0.207	0.23	0.25	0.90	0.90	0.95	0.95
	30	-0.9	-1.59	0.19	0.192	0.231	0.23	0.90	0.90	0.95	0.95
η_2	10	20.7	-1.64	0.068	0.066	0.081	0.081	0.94	0.90	0.98	0.95
	30	21.5	-1.74	0.067	0.061	0.074	0.071	0.94	0.90	0.98	0.95

C.I., confidence interval; $\eta_1 = E(Y)$; $\eta_2 = \text{pr}(Y < 1)$; Rubin/New, Rubin's/New variance estimator.

method. Rubin's method provides good empirical coverage for $\eta_1 = E(Y)$ in the sense that the empirical coverage is close to the nominal coverage; however, the empirical coverage for $\eta_2 = \text{pr}(Y < 1)$ reaches to 95% for 90% confidence intervals, and 98% for 95% confidence intervals, due to variance overestimation. In contrast, our new method provides more accurate coverage of confidence interval for both $\eta_1 = E(Y)$ and $\eta_2 = \text{pr}(Y < 1)$ at 90% and 95% levels.

6. Discussion

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

Another extension would be developing unbiased variance estimation for the vector case of η with $q > 1$ components. As in the scalar case, we can construct the multivariate analogues of the multiple imputation estimator and the variance estimator; however, finding an adequate reference distribution for the statistic $(\hat{\eta}_{\text{MI}} - \eta)^T \hat{V}_{\text{MI}}^{-1} (\hat{\eta}_{\text{MI}} - \eta)/q$ is more subtle in the vector case 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 Xie & Meng (2014) and Li et al. (1994).

ACKNOWLEDGMENT

We are grateful to Xianchao Xie and Xiaoli Meng for many helpful conversations and to the *Biometrika* editors and four referees for their valuable comments that helped to improve this paper. The research of the second author was partially supported by a grant from US National Science Foundation and also by a Cooperative Agreement between the U.S. Department of Agriculture Natural Resources Conservation Service and Iowa State University.

SUPPLEMENTARY MATERIAL

The supplementary material available at *Biometrika* online includes the proof of Theorem 1, the proof of Theorem 2, verification of the new variance estimator being unbiased when $\hat{\eta}_n$ is the maximum likelihood estimator of $\eta = E\{g(Y)\}$, and an 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. *Proc. Surv. Res. Meth. Sect. Am. Statist. Ass.*, 281–6.

BLACKWELL, M., HONAKER, J. & KING, G. (2015). A Unified Approach to Measurement Error and Missing Data: Details and Extensions. *Soc. Meth. Res.*, In press.

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?. *Proc. Surv. Res. Meth. Sect. Am. Statist. Ass.*, 227–32.

FAY, R. E.(1993). Valid inference from imputed survey data. *Proc. Surv. Res. Meth. Sect. Am. Statist. Ass.*, 41–8.

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. *Incomplete data in sample surveys* **2**, 185–207.

HONAKER, J. AND KING, G. & BLACKWELL, M. M.(2010). Package Amelia.

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

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.

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.

KOTT, P. S.(1995). A paradox of multiple imputation. *Proc. Surv. Res. Meth. Sect. Am. Statist. Ass.*, 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**, 685–689.

MENG, X.(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.

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*. John Wiley & Sons.

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.

SATTERTHWAITE, F. E.(1946). An approximate distribution of estimates of variance components. *Biometrics bulletin* **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 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.(2014). Dissecting multiple imputation from a multi-phase inference perspective: what happens when God's, imputer's and analyst's models are uncongenial?. *Statistica Sinica* , to appear.

[*Received XX* $20xx$. *Revised XX* $20xx$]