Investigation of multiple imputation variance estimation

Jae-Kwang Kim[1]

ABSTRACT

Multiple imputation, proposed by Rubin, is a procedure for handling missing data. One of the attractive parts of multiple imputation is the simplicity of the variance estimation formula. Because of the simplicity, it has been often abused and misused beyond its original prescription. This paper provides the bias of the multiple imputation variance estimator for a linear point estimator and discusses when the bias can be safely neglected.

KEY WORDS. *Bayesian; Missing data; Nonresponse; Survey sampling.*

1 Introduction

Multiple imputation is a widely used method of handling missing data in biostatistical and other investigations, including sample surveys. Rubin (1987) provides a comprehensive description of multiple imputation, and Rubin (1996) cites an extensive bibliography on the technique.

Multiple imputation is applied to a data set with missing items by repeating the process of forming a completed data set several, say M, times, thus creating M completed data sets. With multiple imputation, each of the M completed data sets can be used to estimate a population parameter θ, and the overall estimate is the average of these M estimates. The variance of this average is then obtained as the sum of two terms: the average of the variances of the individual estimates from each data set computed in a standard way treating imputed values as observed values; and a term involving the variance between the individual estimates. The variance estimator computed in a standard way treating imputed values as if observed is called *naïve variance estimator.* Let $\hat{\theta}_{I(k)}$ be the imputed estimator of θ based on the k-th imputed data set and let $\hat{V}_{I(k)}$ be the naïve variance estimator of $\hat{\theta}_{I(k)}$. Then, the multiple imputation estimator of θ is

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

and the associated variance estimator is

$$\hat{V}_{M,n} = U_{M,n} + \left(1 + M^{-1}\right) B_{M,n}, \tag{2}$$

where $U_{M,n} = M^{-1} \sum_{k=1}^{M} \hat{V}_{I(k)}$ and $B_{M,n} = (M-1)^{-1} \sum_{k=1}^{M} \left(\hat{\theta}_{I(k)} - \hat{\theta}_{M,n}\right)^2$. We add a subscript for the sample size n to $\hat{\theta}_{M,n}$ and $\hat{V}_{M,n}$ since later asymptotic results are functions

[1] Department of Statistics, Hankuk University of Foreign Studies, Yongin, Kyungki-Do 449-791, Korea.

of M and n. Rubin (1987) suggested using $\hat{\theta}_{M,n}$ as a point estimator of θ and using $\hat{V}_{M,n}$ as a variance estimator of $\hat{\theta}_{M,n}$ We call the variance estimator $\hat{V}_{M,n}$ in (2) *Rubin's variance estimator.*

Multiple imputation often make assumptions about the population to generate imputed values. The assumptions to support the imputation procedure can be described as models, called *population model.* Population model assumes that the finite population is a random sample from an infinite population. The purpose of this paper is to investigate the frequentist validity of the multiple imputation variance estimators under the population model used to create multiple imputation. The starting point for our investigation is to express the multiple imputation point estimator as

$$\hat{\theta}_{M,n} = \hat{\theta}_n + \left(\hat{\theta}_{M,n} - \hat{\theta}_n\right) \tag{3}$$

where $\hat{\theta}_n$ is the full sample estimator without missing data. From (3), the total variance of the multiple imputation point estimator is

$$Var\left(\hat{\theta}_{M,n}\right) = Var\left(\hat{\theta}_n\right) + Var\left(\hat{\theta}_{M,n} - \hat{\theta}_n\right) + 2Cov\left(\hat{\theta}_n, \hat{\theta}_{M,n} - \hat{\theta}_n\right) \tag{4}$$

In this paper, we show that $U_{M,n}$ from (2) estimates $Var\left(\hat{\theta}_n\right)$ for most standard stochastic imputation schemes including a Bayesian imputation scheme. We also show that $B_{M,n}$ + $M^{-1}B_{M,n}$ from (2) estimates $Var\left(\hat{\theta}_{M,n} - \hat{\theta}_n\right)$ when a Bayesian imputation scheme is used. Rubin's variance estimator assumes that the covariance term in (4) is zero. Under the population model approach, we show that Rubin's variance estimator is not always unbiased because the covariance term is sometimes not equal to zero.

For the evaluation of the multiple imputation variance estimator, we consider the joint distribution of the population model (ζ), the sampling mechanism (p), the response mechanism (R), and the imputation mechanism (I). To avoid extraneous issues, we assume that

(A.1) The sampling mechanism, the response mechanism, and the imputation mechanism are ignorable under the assumed population model ζ.

(A.2) The complete sample point estimator is linear in the y-variable and approximately design unbiased for the population mean. The complete sample variance estimator is quadratic in the y-variable and is design unbiased.

(A.3) The imputed and original values have the same expected values:

$$E_\zeta\left(\eta_{i(k)}\right) = E_\zeta\left(Y_i\right), \tag{5}$$

where $\eta_{i(k)}$ is the imputed value associated with unit i for the k-th imputed data set.

(A.4) Let $\eta_{(k)}$ be the k-th imputed data set. Then, the M values of the imputed data set are identically distributed:

$$Pr\left(\eta_{(k)} \in B\right) = Pr\left(\eta_{(l)} \in B\right), \quad \forall k, l \leq M. \tag{6}$$

for any measurable set B.

Under (A.2), condition (5) is a sufficient condition for the imputed estimator to be unbiased. By (6), the M naive variance estimators $\hat{V}_{I(k)}$ are identically distributed.

Given the above assumptions, we examine in the next two sections the conditions under which $\hat{V}_{M,n}$ in (2) is unbiased for $Var\left(\hat{\theta}_{M,n}\right)$ in (4). For this purpose, we consider the two components of $\hat{V}_{M,n}$ separately. Section 2 determines general conditions under which the naïve variance estimator is approximately unbiased for the variance of the estimate based on complete response, $Var\left(\hat{\theta}_n\right)$, and then applies the results to $U_{M,n}$. With a Bayesian imputation scheme, $U_{M,n}$ is shown to be asymptotically unbiased for $Var\left(\theta_n\right)$. Section 3 shows that $\left(1+M^{-1}\right)B_{M,n}$ estimates $Var\left(\hat{\theta}_{M,n}-\hat{\theta}_n\right)$ when an appropriate Bayesian imputation scheme is used. Section 4 considers the overall conditions for $\hat{V}_{M,n}$ to be approximately unbiased for $Var\left(\hat{\theta}_{M,n}\right)$ and provides some concluding remarks. The proofs of the theorems are not provided here for brevity.

2 Evaluation of the within-imputation variance component

This section examines the use $U_{M,n}$ to estimate $Var\left(\hat{\theta}_n\right)$. With $U_{M,n}=M^{-1}\sum_{k=1}^{M}\hat{V}_{I(k)}$, and the fact that the $\hat{V}_{I(k)}$ are identically distributed under assumption (A.4), examining the unbiasedness of $U_{M,n}$ is equivalent to examining the unbiasedness of the naïve variance estimator $\hat{V}_{I(k)}$ for any one of the replicate data set.

We first establish a general lemma concerning the bias of the naïve variance estimator $\hat{V}_I$ for estimating $Var\left(\hat{\theta}_n\right)$ and a theorem that gives the condition for the naïve variance estimator to be asymptotically unbiased for $Var\left(\hat{\theta}_n\right)$. These results are applicable for any form of imputation, including Bayesian imputation.

To derive the general lemma, we note that the complete sample variance estimator is a quadratic function of the sample values, and can therefore in general be written as

$$\hat{V}_n=\sum_{i\in A}\sum_{j\in A}\Omega_{ij}Y_iY_j \tag{7}$$

for some coefficients Ω_{ij}, where A is the set of indices for the sample. Thus, the expectation of $\hat{V}_I$ is equal to the expectation of $\hat{V}_n$ if the mean and the covariance structures of the imputed data set are the same as those of the original data set. Although the mean structures are the same from assumption (A.3), the covariance structure need not be the same.

The following general result expresses the bias of the naïve variance estimator for estimating the anticipated sampling variance as a function of the difference of the covariance term between the imputed data and the original data.

Lemma 2.1 *Assume (A.1)-(A.3) hold. Then, the bias of the naïve variance estimator as an estimator of the sampling variance is*

$$E\left(\hat{V}_I\right)-Var\left(\hat{\theta}_n\right)=E\left(\sum_{i\in A}\sum_{j\in A}\Omega_{ij}\tau_{ij}\right) \tag{8}$$

where Ω_{ij} are the coefficients in (7) and $\tau_{ij} = Cov_{\zeta I}(\eta_i, \eta_j) - Cov_{\zeta}(Y_i, Y_j)$ is the difference between the covariance of the imputed values and the covariance of the original values.

The following theorem gives conditions under which the bias term is negligible.

Theorem 2.1 *Assume (A.1)-(A.3) hold. Assume a sequence of finite populations as described in Isaki and Fuller (1981) with finite fourth moments. Assume that*

$$\sum_{i\in A}\sum_{j\in A} |\Omega_{ij}| = O\left(n^{-1}\right). \tag{9}$$

If

$$\max_{i,j} \tau_{ij} = o_p(1), \tag{10}$$

then

$$\lim_{n\to\infty} n\left\{E\left(\hat{V}_I\right) - Var\left(\hat{\theta}_n\right)\right\} = 0, \tag{11}$$

where the Ω_{ij} are the coefficient used in (7) and $\tau_{ij} = Cov_{\zeta I}(\eta_i, \eta_j) - Cov_{\zeta}(Y_i, Y_j)$ is the difference of the covariance between the imputed values and the original values.

The theorem follows directly from Lemma 2.1. Condition (9) generally holds for many sampling designs, including stratified cluster sampling designs. Condition (10) requires that the covariance structure for the imputed values is asymptotically the same as that for the original values. This condition holds for many random imputation schemes, when the number of respondents is much larger than the number of parameters in the imputation model. Condition (10) also holds with Bayesian imputation, if the posterior values of the missing items are created with sufficient degrees of freedom. For more details, see Kim (2002).

3 Evaluation with the between-imputation variance component

The section evaluates the between-imputation variance component $\left(1 + M^{-1}\right) B_{M,n}$ in (2). To do this, we express $\hat{\theta}_{M,n}$ as

$$\hat{\theta}_{M,n} = \hat{\theta}_n + \left(\hat{\theta}_{\infty,n} - \hat{\theta}_n\right) + \left(\hat{\theta}_{M,n} - \hat{\theta}_{\infty,n}\right) \tag{12}$$

where $\hat{\theta}_{\infty,n} = \lim_{M\to\infty} \hat{\theta}_{M,n}$ is the infinite-M multiple imputation point estimator. The previous section has considered the use of $U_{M,n}$ to estimate the variance of $\hat{\theta}_n$. We are now concerned with the second and third terms on the right hand side of (12).

The following theorem shows that, in general, $B_{M,n}$ is unbiased for the variance of the second term on the right hand side of (12) if Bayesian imputation is used.

Theorem 3.1 *Let Y_{obs} and Y_{mis} be the observed part and the missing part of the sample, respectively. Assume Bayesian imputation, where the imputed values are independently drawn from the conditional distribution $L(Y_{mis} \mid Y_{obs})$ of Y_{mis} given Y_{obs}. Assume that there exists a positive δ such that*

$$\int \hat{\theta}_n\left(Y_{obs}, Y_{mis}\right)^{2+\delta} dL\left(Y_{mis} \mid Y_{obs}\right) < \infty \tag{13}$$

almost everywhere in Y_{obs}. Then, we have

$$\lim_{M\to\infty}\left\{E\left(B_{M,n}\right)-Var\left(\hat{\theta}_{M,n}-\hat{\theta}_{\infty,n}\right)\right\}=0 \tag{14}$$

for all n.

We now consider the case of finite M, that is, we consider the third term, $\left(\hat{\theta}_{M,n}-\hat{\theta}_{\infty,n}\right)$, in (12). In the following theorem, we show that $M^{-1}B_{M,n}$ can be used to estimate the variance of $\left(\hat{\theta}_{M,n}-\hat{\theta}_{\infty,n}\right)$ and that the covariance between $\hat{\theta}_{\infty,n}$ and $\left(\hat{\theta}_{M,n}-\hat{\theta}_{\infty,n}\right)$ is equal to zero.

Theorem 3.2 *Assume (A.4) holds so that the M imputed estimators, $\hat{\theta}_{I(k)}$, are identically distributed. Then, we have*

$$E\left(M^{-1}B_{M,n}\right)=Var\left(\hat{\theta}_{M,n}-\hat{\theta}_{\infty,n}\right) \tag{15}$$

and

$$Cov\left(\hat{\theta}_{\infty,n},\hat{\theta}_{M,n}-\hat{\theta}_{\infty,n}\right)=0 \tag{16}$$

for all $M>2$ and n.

4 Conclusion

By combining the results of Lemma 2.1, Theorem 3.1, and Theorem 3.2, we obtain the following result.

Theorem 4.1 *Assume (A.1)-(A.4) and the assumptions of Theorem 3.1 hold. Then, the bias of Rubin's variance estimator is, for large n and M,*

$$Bias\left(\hat{V}_{M,n}\right)\doteq E\left(\sum_{i\in A}\sum_{j\in A}\Omega_{ij}\tau_{ij}\right)-2Cov\left(\hat{\theta}_n,\hat{\theta}_{M,n}-\hat{\theta}_n\right), \tag{17}$$

where Ω_{ij} are the coefficients in (7) and $\tau_{ij}=Cov_{\zeta I}\left(\eta_i,\eta_j\right)-Cov_{\zeta}\left(Y_i,Y_j\right)$ is the difference between the covariance of the imputed values and the covariance of the original values.

The first term in the right side of (17) is the bias of $U_{M,n}$ as an estimator of the variance of $\hat{\theta}_n$ and will be negligible for many random imputations. The second term is a potential bias that is not captured by Rubin's variance estimator. Note that, since $\hat{\theta}_{I(k)}$ $(k=1,2,\cdots,M)$ are identically distributed, $Cov\left(\hat{\theta}_n,\hat{\theta}_{M,n}-\hat{\theta}_n\right)=Cov\left(\hat{\theta}_n,\hat{\theta}_{I(k)}-\hat{\theta}_n\right)$. Hence, this second term is not dependent on M.

A sufficient condition for the second term to be negligible is

$$E_R\left(\hat{\theta}_{M,n}\right)-\hat{\theta}_n=o_p\left(n^{-1/2}\right). \tag{18}$$

To see this, note that

$$\begin{aligned} nCov_{\zeta pR}\left(\hat{\theta}_n,\hat{\theta}_{M,n}-\hat{\theta}_n\right) &= Cov_{\zeta p}\left\{n^{1/2}\hat{\theta}_n,n^{1/2}E_R\left(\hat{\theta}_{M,n}-\hat{\theta}_n\right)\right\} \\ &\quad +nE_{\zeta p}\left\{Cov_R\left(\hat{\theta}_n,\hat{\theta}_{M,n}-\hat{\theta}_n\right)\right\}. \end{aligned}$$

The first term in the right side of the above equality is $o(1)$ as by (18) and the second term is also equal to zero because $\hat{\theta}_n$ is a constant under the response mechanism.

Assumption (18) implies a negligible correlation between the sampling error and the imputation error. In Rubin (1987) this condition is the first requirement of proper imputation. Note that assumption (18) is not an essential requirement for the approximate unbiasedness of the multiple imputation point estimator. The point estimator is in fact approximately unbiased under the weaker assumption that

$$E_{\zeta p}\left\{E_R\left(\hat{\theta}_{M,n}\right)-\hat{\theta}_n\right\}=o\left(n^{-1/2}\right). \tag{19}$$

In particular, domain estimation with the iid model satisfies (19) but not (18). However, if only (19) holds, we have to estimate the covariance term to obtain an unbiased variance estimate rather than rely on Rubin's variance estimator.

Although not derived here, we have also investigated the covariance term in (4) under the linear regression models used by Schenker and Welsh (1988), and concluded that the covariance can be made to equal zero by the inclusion of the appropriate set of auxiliary variables in the model. In particular, when the complete sample estimator is linear in the y-variable, then Rubin's variance estimator is unbiased for the variance of that estimator provided that the final weight used in producing the estimator is included as one of auxiliary variables in the imputation model. The final weight for estimators of parameters of the total population is the standard weight in the data file. However, for a subgroup estimator, the final weight is the weight in the data file for unit in the subgroup and is zero for other units. For preplanned subgroup analyses, it may be possible to include all the required sets of final weights in the model. However, for unplanned subgroup analysis this will not be possible. If the final weights for a particular estimator are not included in the model, then Rubin's variance estimator will be biased. The domain estimation problem highlighted by Fay (1992) is of this type.

REFERENCES

Fay, R.E. (1992) "When are inferences from multiple imputation valid ?" *ASA Proceedings of the Section on Survey Research Methods,* 227-232 .

Isaki, C. and Fuller, W. A. (1982) "Survey design under the regression superpopulation model." *Journal of the American Statistical Association,* 77, 89-96.

Kim, J. K. (2002) "A note on approximate Bayesian bootstrap imputation." *Biometrika.* In press.

Rubin, D.B. (1987). *Multiple imputation for nonresponse in surveys.* New York: Wiley.

Rubin, D.B. (1996) "Multiple Imputation After 18+ years." *Journal of the American Statistical Association,* 91, 473-489.

Schenker, N. and Welsh, A. H. (1988). "Asymptotic results for multiple imputation". *The Annals of Statistics,* 16, 1550-1566