2006

# Imputation using response probability

Jae Kwang Kim, *Yonsei University*
Hyeonah Park, *Seoul National University*

*The Canadian Journal of Statistics*
Vol. 34, No. 1, 2006, Pages 1–12
*La revue canadienne de statistique*

# Imputation using response probability

Jae Kwang KIM and Hyeonah PARK

*Key words and phrases:* Nonresponse; survey sampling; variance estimation.

*MSC 2000:* Primary 62D05; secondary 62G09.

*Abstract:* The authors propose a new ratio imputation method using response probability. Their estimator can be justified either under the response model or under the imputation model; it is thus doubly protected against the failure of either of these models. The authors also propose a variance estimator that can be justified under the two models. Their methodology is applicable whether the response probabilities are estimated or known. A small simulation study illustrates their technique.

## Utilisation de probabilités de réponse à des fins d'imputation

*Résumé :* Les auteurs proposent une nouvelle méthode d'imputation par quotient basée sur les probabilités de réponse. Leur estimateur peut être justifié au moyen du modèle de réponse et du modèle d'imputation ; il jouit ainsi d'une double protection en cas d'invalidité de l'un ou l'autre de ces modèles. Les auteurs proposent aussi un estimateur de variance justifiable sous les deux modèles. Leur méthodologie peut être utilisée tant lorsque les probabilités de réponse sont estimées que connues. Une petite étude de simulation illustre leur propos.

## 1. INTRODUCTION

Imputation is a commonly used method of compensating for item nonresponse in sample surveys. Reasons for conducting imputation are to facilitate analyses using complete data analysis methods, to ensure that the results obtained by different analyses are consistent with one another, and to reduce nonresponse bias. Kalton (1983) and Groves, Dillman, Eltinge & Little (2002) provide a comprehensive overview of imputation methods in survey sampling.

Many imputation methods such as ratio imputation or regression imputation use auxiliary information that is observed throughout the sample. Such imputation methods require assumptions about the distribution of the study variable. The imputation model refers to the assumptions about the variables collected in the survey and the relationship among these variables. Another model, called the response model, is also commonly adopted in the analysis of missing data. The response model refers to the assumptions about the probability of obtaining responses from the sample for the item. One of the commonly used response models is the uniform response model, where the responses are assumed to be independent and identically distributed within the imputation cell. Rao & Shao (1992), Rao & Sitter (1995) and Shao & Steel (1999) discuss inference using the imputed estimator under the uniform response model. However, for the other nonuniform response models such as the logistic response model, imputation methods incorporating the response model are relatively underdeveloped, although analyses incorporating the response model are quite popular in the nonimputation context. Examples include Rosenbaum (1987), Robins, Rotnitzky & Zhao (1994), and Lipsitz, Ibrahim & Zhao (1999).

In this article, we provide an imputation methodology that combines the imputation model and the response model. The proposed method can be justified under either one of the two approaches. That is, it is justified if either a response model or an imputation model can be correctly specified. Thus, the resulting estimator is doubly protected against the failure of the assumed model. (Scharfstein, Rotnitsky & Robins 1999). The basic project is introduced under the ratio imputation model in Section 2. The proposed method is further discussed in Section 3. In Section 4, we propose a replication variance estimator that can be justified under the two models. In Section 5, we discuss the proposed imputation method when the response probabilities are estimated rather than known. A limited simulation study is presented in Section 6. Concluding remarks are made in Section 7.

## 2. BASIC SETUP

Let the finite population be of size $N$, indexed from 1 to $N$. Let the parameter of interest be the population total $Y = y_1 + \cdots + y_N$, where $y_i$ is the study variable of unit $i$. Let $\mathcal{F} = \{y_1, \ldots, y_N\}$ be the collection of the study variable in the finite population. Let $\widehat{Y}_n$ be an estimator of the population total $Y$ based on the sample of size $n$ and of the form $\widehat{Y}_n = \sum_{i \in A} w_i y_i$, where $w_i$ is the sampling weight of unit $i$ and $A$ is the set of indices in the sample. We assume that

$$\mathrm{E}\,(\widehat{Y}_n \,|\, \mathcal{F}) = Y, \tag{1}$$

where the expectation is taken with respect to the sampling mechanism. Condition (1), which can be relaxed later, clearly means that $\widehat{Y}_n$ is the Horvitz–Thompson estimator.

Under nonresponse, we define the response indicator function of $y_i$

$$R_i = \begin{cases} 1 & \text{if } y_i \text{ responds,} \\ 0 & \text{if } y_i \text{ does not respond,} \end{cases}$$

and let $\mathcal{R} = \big\{(i, R_i) : i \in A\big\}$ be the set of response indicators for all units in the sample. If we define $y_i^*$ to be the imputed value of $y_i$, then the estimator of $Y$ based on the imputed values can be written

$$\widehat{Y}_I = \sum_{i \in A} w_i \big\{R_i y_i + (1 - R_i) y_i^*\big\}. \tag{2}$$

Suppose that we have an auxiliary variable $x_i$ for unit $i$ in the sample and that the $x_i$ are completely observed throughout the sample. We assume that

$$\mathrm{E}\,(y_i \,|\, A, \mathcal{R}, \mathcal{X}) = x_i \gamma, \quad i = 1, \ldots, N \tag{3}$$

where $\mathcal{X} = \big\{(i, x_i) : i \in A\big\}$ and the expectation in (3) is with respect to the conditional distribution of $y_i$ given the realized sample, the realized $x$-values, and the realized respondent status. Under the ratio model in (3), the imputed value of $y_i$ takes the form of $y_i^* = x_i \hat{\gamma}$, where $\hat{\gamma}$ is to be determined. Often, for example in Rao (1996) and in Rao & Sitter (1995), the choice of $\hat{\gamma}$ was

$$\hat{\gamma}_R = \Big\{\sum_{i \in A} w_i R_i x_i\Big\}^{-1} \sum_{i \in A} w_i R_i y_i. \tag{4}$$

The ratio imputation using $y_i^* = x_i \hat{\gamma}_R$ satisfies $\mathrm{E}\,(\widehat{Y}_I - \widehat{Y}_n \,|\, A, \mathcal{X}, \mathcal{R}) = 0$ by (3), and so by (1), $\mathrm{E}\,(\widehat{Y}_I - Y) = 0$. The unbiasedness of the imputed estimator in (2) depends on assumption (3). If assumption (3) fails, then we cannot guarantee the unbiasedness of the imputed estimator.

Let $\pi_i = \mathrm{P}\,(R_i = 1 \,|\, i \in A)$ be the response probability of sampled unit $i$. If we know the response probability $\pi_i$, then we can use the information about response probability to relax assumption (3). The proposed estimator is

$$\widehat{Y}_{\mathrm{Id}} \;=\; \sum_{i \in A} w_i y_i^* + \sum_{i \in A} w_i \pi_i^{-1} R_i (y_i - y_i^*). \tag{5}$$

Note that expression (5) is essentially the two-phase sampling estimator, where $R_i$ corresponds to the second-phase sampling indicator and $y_i^*$ is the predicted value of $y_i$ using the second-phase sample observation and the first-phase auxiliary information.

Note that the estimator in (5) can be written as that in (2) if and only if

$$\sum_{i \in A} w_i (\pi_i^{-1} - 1) R_i (y_i - y_i^*) = 0. \tag{6}$$

Hence, a choice of $\hat{\gamma}$ satisfying (6) is

$$\hat{\gamma}_{M2} = \left\{ \sum_{i \in A} w_i (\pi_i^{-1} - 1) R_i x_i \right\}^{-1} \sum_{i \in A} w_i (\pi_i^{-1} - 1) R_i y_i, \tag{7}$$

which reduces to (4) under the uniform response model, since the $\pi_i$ are all equal. The imputed estimator $\widehat{Y}_I$ using $\hat{\gamma}_{M2}$ in (7) is algebraically equivalent to $\widehat{Y}_{\text{Id}}$ in (5). We use $\widehat{Y}_I$ to denote the imputed estimator using $\hat{\gamma}_R$ in (4), and $\widehat{Y}_{\text{Id}}$ to denote the newly proposed estimator using $\hat{\gamma}_{M2}$ in (7).

To discuss the nature of the proposed ratio imputation estimator using $\hat{\gamma}_{M2}$ in (7), we adopt the extended definition of $R_i$ introduced by Fay (1991). Conceptually, the response indicator function $R_i$ can be extended to the entire population. That is, let $R_i$ take the value one if unit $i$ responds when sampled, and the value zero otherwise. Define

$$(X_R, Y_R) = \sum_{i=1}^{N} R_i (x_i, y_i)$$

to be the population total of the conceptual respondents and define

$$(X_M, Y_M) = \sum_{i=1}^{N} (1 - R_i)(x_i, y_i)$$

to be the population total of the conceptual nonrespondents. Conditional on $R_1, \ldots, R_N$, the population respondent total $(X_R, Y_R)$ can be unbiasedly estimated by

$$(\widehat{X}_R, \widehat{Y}_R) = \sum_{i \in A} w_i R_i (x_i, y_i),$$

and the population nonrespondent total $(X_M, Y_M)$ can be unbiasedly estimated by

$$(\widehat{X}_M, \widehat{Y}_M) = \sum_{i \in A} w_i (1 - R_i)(x_i, y_i) \quad \text{or} \quad (\widehat{X}_{M2}, \widehat{Y}_{M2}) = \sum_{i \in A} w_i R_i (\pi_i^{-1} - 1)(x_i, y_i).$$

Using this notation, the two imputed estimators can be written

$$\widehat{Y}_I = \widehat{X}_R \hat{\gamma}_R + \widehat{X}_M \hat{\gamma}_R$$

and

$$\widehat{Y}_{\text{Id}} = \widehat{X}_R \hat{\gamma}_R + \widehat{X}_M \hat{\gamma}_{M2}, \tag{8}$$

where $\hat{\gamma}_R = \widehat{X}_R^{-1} \widehat{Y}_R$ and $\hat{\gamma}_{M2} = \widehat{X}_{M2}^{-1} \widehat{Y}_{M2}$. Since $\hat{\gamma}_R$ estimates the population characteristic of the respondents and $\hat{\gamma}_{M2}$ estimates the population characteristic of the nonrespondents, it makes more sense to use $\hat{\gamma}_{M2}$ to impute for the nonrespondents.

## 3. ASYMPTOTIC PROPERTIES

To discuss the asymptotic properties of $\widehat{Y}_{\text{Id}}$, let us assume a sequence of finite populations with finite fourth moments of $(x_i, y_i)$ as defined in Isaki & Fuller (1982). Assume the sampling mechanism satisfies

$$K_1 < \max_{1 \le i \le N} (n w_i / N) < K_2 \tag{9}$$

and

$$K_3 < n \, \text{var}(\widehat{Y}_n \mid \mathcal{F}) / N^2 < K_4 \tag{10}$$

for some nonnegative constants $K_1$, $K_2$, $K_3$, and $K_4$, uniformly in $n$. Assume that the response mechanism satisfies

$$K_5 < \pi_i, \tag{11}$$

for some nonnegative constant $K_5$, and

$$\mathrm{P}\,(R_i = 1, R_j = 1) = \mathrm{P}\,(R_i = 1)\,\mathrm{P}\,(R_j = 1), \quad \forall\, i \neq j. \tag{12}$$

Let

$$(X^*, Y^*) = \sum_{i=1}^{N} \pi_i^{-1} R_i (x_i, y_i),$$

the probability limit of

$$\sum_{i \in A} w_i \pi_i^{-1} R_i (x_i, y_i)$$

conditional on $R_1, \ldots, R_N$. Under the response model of $\mathrm{E}\,(\pi_i^{-1} R_i) = 1$, $(X^*, Y^*)$ is consistent for $(X, Y)$, the population total of $(x_i, y_i)$. Here, we do not necessarily assume that the response model is true. By a Taylor expansion, it is shown in the Appendix that

$$\widehat{Y}_{\mathrm{Id}} - \widehat{Y}_n = \sum_{i \in A} w_i \big\{\delta(\pi_i^{-1} - 1)R_i - (1 - R_i)\big\}(y_i - \gamma_M^* x_i) + o_p(n^{-1/2}N), \tag{13}$$

where $\delta = (X^* - X_R)^{-1} X_M$, $\gamma_M^* = (X^* - X_R)^{-1}(Y^* - Y_R)$ and $Z_n = o_p(a_n)$ denotes that $a_n^{-1} Z_n$ converges to zero in probability.

Note that if $\mathrm{E}\,(\pi_i^{-1} R_i) = 1$, then $\delta \doteq 1$ and the right-hand side of (13) is asymptotically negligible. Thus, the proposed estimator $\widehat{Y}_{\mathrm{Id}}$ is approximately unbiased for $Y$ under the assumed response model, regardless of whether the ratio imputation model holds or not. On the other hand, if the response model is not correctly specified, we still have approximate unbiasedness under the ratio imputation model (3) because $\mathrm{E}\,(y_i - \gamma_M^* x_i \mid A, \mathcal{X}, \mathcal{R}) = 0$. Hence, the estimator $\widehat{Y}_{\mathrm{Id}}$ is doubly protected since that we require that only one of the two models, a model for the value of $y$ or a model for the response probability, be correctly specified. Lipsitz, Ibrahim & Zhao (1999) made a similar argument for double protection in the problem of constructing a doubly protected estimating equation.

Under the response model, by (1) and by the definition of $\pi_i$,

$$\mathrm{E}\,(\widehat{Y}_{\mathrm{Id}} \mid \mathcal{F}) = Y + o(n^{-1/2}N) \tag{14}$$

and, by (12),

$$\mathrm{var}(\widehat{Y}_{\mathrm{Id}} \mid \mathcal{F}) = \mathrm{var}(\widehat{Y}_n \mid \mathcal{F}) + \mathrm{E}\left\{\sum_{i \in A} w_i^2 (\pi_i^{-1} - 1)(y_i - x_i \gamma_M)^2 \mid \mathcal{F}\right\} + o(n^{-1}N^2), \tag{15}$$

where $\gamma_M = X_M^{-1} Y_M$.

The model expectation of the variance in (15) is minimized when $\mathrm{E}\,(\gamma_M \mid R_1, \ldots, R_N) = \gamma$, where $\gamma$ is the imputation model parameter defined in (3). Thus, the role of the imputation model in the response model approach is to reduce the variance. If the imputation model does not hold, then by (14), the point estimator is still approximately unbiased, but the variance will be generally large. This is consistent with the general philosophy of model-assisted estimation in survey sampling: If the model is good, then the point estimator is efficient. Otherwise, the point estimator is still approximately design unbiased (Särndal, Swensson & Wretman 1992).

Under the ratio imputation model defined by (3) and

$$\mathrm{cov}(y_i, y_j \mid A, \mathcal{R}, \mathcal{X}) = \begin{cases} \sigma_i^2 & \text{if } i = j, \\ 0 & \text{if } i \neq j, \end{cases} \tag{16}$$

it can be shown that $\mathrm{E}\,(\widehat{Y}_{\mathrm{Id}} - Y) = o(n^{-1/2}N)$, and under $N^{-1}n = o(1)$,

$$\mathrm{var}(\widehat{Y}_{\mathrm{Id}} - Y) = \mathrm{var}\left(\sum_{i\in A} w_i x_i \gamma \,|\, \mathcal{F}\right) + \mathrm{E}\left(\sum_{i\in A} w_i^2 \pi_i^{-2} q_i^2 R_i \sigma_i^2 \,|\, \mathcal{F}\right) + o(n^{-1}N^2), \qquad (17)$$

where $q_i = \delta(1-\pi_i) + \pi_i$.

Inserting $\pi_i^{-1} - 1$ into the estimation of the ratio coefficient $\gamma$ was first proposed by Brewer (1979) for the complete sample, where the $\pi_i$ in this case is the inclusion probability of unit $i$. Thus, the estimator using (7) is the extension of Brewer's idea to the imputation problem.

## 4. VARIANCE ESTIMATION

We now consider the variance estimation of the proposed ratio imputation estimator satisfying (6). We adopt a replication method such as the jackknife for variance estimation. Replication variance estimation is popular because it can be easily extended to variance estimation for nonlinear statistics.

Under complete response, let a replication variance estimator be

$$\widehat{V}_n = \sum_{k=1}^{L} c_k (\widehat{Y}_n^{(k)} - \widehat{Y}_n)^2, \qquad (18)$$

where $\widehat{Y}_n^{(k)}$ is the $k$th estimate of $Y$ based on the observations included in the $k$th replicate, $L$ is the number of replicates, and $c_k$ is a factor associated with replicate $k$ determined by the replication method. When the original estimator $\widehat{Y}_n$ is a linear estimator, the $k$th replicate of $\widehat{Y}_n$ can be written

$$\widehat{Y}_n^{(k)} = \sum_{i\in A} w_i^{(k)} y_i,$$

where $w_i^{(k)}$ denotes the replicate weight for the $i$th unit of the $k$th replication. We assume that

$$\mathrm{E}\left[\{\widehat{V}_n / \mathrm{var}(\widehat{Y}_n) - 1\}^2 \,|\, \mathcal{F}\right] \quad = \quad o(1) \qquad (19)$$

for any $y$ with bounded fourth moments.

Under nonresponse, we propose a variance estimator for the imputed estimator of the form in (5) using the replication method in (18). The proposed replication variance estimator is

$$\widehat{V}_d = \sum_{k=1}^{L} c_k (\widehat{Y}_{\mathrm{Id}}^{(k)} - \widehat{Y}_{\mathrm{Id}})^2,$$

where

$$\widehat{Y}_{\mathrm{Id}}^{(k)} = \sum_{i\in A} w_i^{(k)} y_i^{*(k)} + \sum_{i\in A} w_i^{(k)} \pi_i^{-1} R_i (y_i - y_i^{*(k)})$$

and $y_i^{*(k)}$ is the $k$th replicate of $y_i^*$ satisfying

$$\sum_{i\in A} w_i^{(k)} (\pi_i^{-1} - 1) R_i (y_i - y_i^{*(k)}) = 0. \qquad (20)$$

Note that condition (20) for the replicates is similar to condition (6) for the original estimator.

For the ratio imputation $y_i^* = x_i \hat{\gamma}_{M2}$, the replicate is $y_i^{*(k)} = x_i \hat{\gamma}_{M2}^{(k)}$, where

$$\hat{\gamma}_{M2}^{(k)} = \left\{\sum_{i\in A} w_i^{(k)} (\pi_i^{-1} - 1) R_i x_i\right\}^{-1} \sum_{i\in A} w_i^{(k)} (\pi_i^{-1} - 1) R_i y_i.$$

Note that (20) implies

$$\widehat{Y}_{\mathrm{Id}}^{(k)} = \sum_{i \in A} w_i^{(k)} \{R_i y_i + (1 - R_i) y_i^{*(k)}\}. \tag{21}$$

If we have a uniform response mechanism across the whole sample, then the replicates in (21) reduce to those of the adjusted jackknife method of Rao (1996).

To show the consistency of $\widehat{V}_d$ to $\mathrm{var}(\widehat{Y}_{\mathrm{Id}} \,|\, \mathcal{F})$, we assume, in addition to (9)–(12),

$$\mathrm{E}\left[\{c_k(\widehat{Y}_n^{(k)} - \widehat{Y}_n)^2\}^2 \,|\, \mathcal{F}, R_1, \ldots, R_N\right] < C_y L^{-2} \big\{\mathrm{var}(\widehat{Y}_n \,|\, \mathcal{F}, R_1, \ldots, R_N)\big\}^2 \tag{22}$$

for all $k$ and for some constant $C_y$. Then, by a Taylor expansion, it is shown in the Appendix that

$$c_k^{1/2}(\widehat{Y}_{\mathrm{Id}}^{(k)} - \widehat{Y}_{\mathrm{Id}}) = c_k^{1/2} \sum_{i \in A} (w_i^{(k)} - w_i)\zeta_i + o_p(L^{-1/2} n^{-1/2} N), \tag{23}$$

where $\zeta_i = x_i \gamma_M^* + \pi_i^{-1} R_i q_i (y_i - x_i \gamma_M^*)$ with $q_i = \delta(1 - \pi_i) + \pi_i$. Thus, by (19),

$$\sum_{k=1}^{L} c_k (\widehat{Y}_{\mathrm{Id}}^{(k)} - \widehat{Y}_{\mathrm{Id}})^2 = \mathrm{var}\left(\sum_{i \in A} w_i \zeta_i \,|\, \mathcal{F}, R_1, \ldots, R_N\right) + o_p(n^{-1} N^2) \tag{24}$$

and note that, by (13), $\sum_{i \in A} w_i \zeta_i = \widehat{Y}_{\mathrm{Id}} + o_p(n^{-1/2} N)$ and the proposed variance estimator is consistent for the conditional variance of $\widehat{Y}_{\mathrm{Id}}$ conditional on $R_1, \ldots, R_N$. To show the consistency for the unconditional variance, note that

$$\mathrm{E}\left(\sum_{i \in A} w_i \zeta_i \,|\, \mathcal{F}, R_1, \ldots, R_N\right) = \sum_{i=1}^{N} \zeta_i \quad \text{and} \quad \mathrm{var}\left(\sum_{i=1}^{N} \zeta_i \,|\, \mathcal{F}\right) = O(N),$$

which is of a smaller order than that of the conditional variance when $N^{-1} n = o(1)$. Finally, it remains to show that $\mathrm{var}\big(\sum_{i \in A} w_i \zeta_i \,|\, \mathcal{F}, R_1, \ldots, R_N\big)$ is consistent for $\mathrm{E}\big\{\mathrm{var}\big(\sum_{i \in A} w_i \zeta_i \,|\, \mathcal{F}, R_1, \ldots, R_N\big) \,|\, \mathcal{F}\big\}$. The argument for the consistency of the conditional variance to its expectation is the same as that of Kim, Navarro & Fuller (2006, Eq. B.11) and is not discussed here.

Note that (24) is derived conditional on $R_1, \ldots, R_N$ and $\mathcal{F}$, and is thus valid without using the response model or the imputation model. Under the response model, it can be shown that

$$(\gamma_M^*, \delta) = (\gamma_M, 1) + O(N^{-1/2}).$$

Thus, we have

$$\begin{aligned} &\mathrm{E}\left\{\mathrm{var}\left(\sum_{i \in A} w_i \zeta_i \,|\, \mathcal{F}, R_1, \ldots, R_N\right) \,\Big|\, \mathcal{F}\right\} \qquad\qquad (25) \\ &= \quad \mathrm{var}(\widehat{Y}_n \,|\, \mathcal{F}) + E\left\{\sum_{i \in A} w_i^2 (\pi_i^{-1} - 1)(y_i - x_i \gamma_M)^2 \,|\, \mathcal{F}\right\} + o(n^{-1} N^2), \end{aligned}$$

which is equal to (15), proving the consistency under the response model, where the expectation in (25) is over the response model. Under the ratio imputation model, assuming $\max_i \sigma_i^2 = O(1)$, it can be shown that

$$\gamma_M^* = \gamma + O(N^{-1/2})$$

and so

$$\begin{aligned} &\mathrm{E}\left\{\mathrm{var}\left(\sum_{i \in A} w_i \zeta_i \,|\, \mathcal{F}, R_1, \ldots, R_N\right) \,\Big|\, R_1, \ldots, R_N\right\} \\ &= \quad \mathrm{var}\left(\sum_{i \in A} w_i x_i \gamma \,|\, \mathcal{F}\right) + \mathrm{E}\left\{\sum_{i \in A} w_i^2 \pi_i^{-2} q_i^2 R_i \sigma_i^2 \,|\, \mathcal{F}\right\} + o(n^{-1} N^2), \end{aligned}$$

which is equalto (17), proving the consistency under the ratio imputation model, where the expectation in (25) is over the ratio imputation model defined in (3) and (16).

Note that the variance estimator discussed above is consistent only when the sampling fraction $f = n/N$ is negligible. If such a condition is not satisfied, then a consistent estimator of $\operatorname{var}(\zeta_1 + \cdots + \zeta_N \mid \mathcal{F})$ should be added to the final variance estimator, as discussed in Shao & Steel (1999). For the ratio imputation described in Section 3, a consistent estimator of $\operatorname{var}(\zeta_1 + \cdots + \zeta_N \mid \mathcal{F})$ is

$$\sum_{i \in A} w_i \pi_i^{-1}(\pi_i^{-1} - 1)\hat{q}_i^2 R_i (y_i - x_i \hat{\gamma}_{M2})^2,$$

where $\hat{\gamma}_{M2}$ is defined following (8) and $\hat{q}_i = \hat{\delta}(1 - \pi_i) + \pi_i$ with $\hat{\delta}$ defined following (A.11) in the Appendix.

## 5. IMPUTATION USING ESTIMATED RESPONSE PROBABILITY

So far, we have assumed that the response probabilities are known. In practice, the response probabilities are unknown and we have to estimate them. In addition to (11) and (12), assume that the model for response probability is a parametric model such that $\pi_i \equiv \pi(x_i; \alpha)$ is a smooth function of $x_i$ and a finite-dimensional parameter $\alpha$ whose range lies in $(0, 1]$. Let $\hat{\pi}_i = \pi(x_i; \hat{\alpha})$ be the estimated response probability of $\pi_i$, where $\hat{\alpha}$ satisfies

$$n^{1/2}(\hat{\alpha} - \alpha) = n^{-1/2} \sum_{i \in A} H(R_i; \alpha) + o_p(1) \tag{26}$$

where

$$\mathrm{E}\{H(R_i; \alpha)\} = 0 \quad \text{and} \quad \mathrm{E}\{H(R_i; \alpha) H(R_i; \alpha)^\top\}$$

is positive definite. For example, the logistic regression model defined by

$$\pi_i = \{1 + \exp(-\alpha_0 - \alpha_1 x_i)\}^{-1}$$

satisfies

$$H(R_i; \alpha) = n\{I(\alpha_0, \alpha_1)\}^{-1}(R_i - \pi_i)(1, x_i)^\top$$

and

$$I(\alpha_0, \alpha_1) = \mathrm{E}\left\{\sum_{i \in A} \pi_i (1 - \pi_i)(1, x_i)^\top (1, x_i)\right\}.$$

The proposed imputation estimator using the estimated response probability is

$$\begin{aligned} \widehat{Y}_{\mathrm{Ie}} &= \sum_{i \in A} w_i \{R_i y_i + (1 - R_i) x_i \hat{\gamma}_{\mathrm{Me}} t\} \\ &= \sum_{i \in A} w_i x_i \hat{\gamma}_{\mathrm{Me}} + \sum_{i \in A} w_i \hat{\pi}_i^{-1} R_i (y_i - x_i \hat{\gamma}_{\mathrm{Me}}), \end{aligned}$$

where

$$\hat{\gamma}_{\mathrm{Me}} = \left\{\sum_{i \in A} w_i (\hat{\pi}_i^{-1} - 1) R_i x_i\right\}^{-1} \sum_{i \in A} w_i (\hat{\pi}_i^{-1} - 1) R_i y_i. \tag{27}$$

As for the asymptotic properties of $\widehat{Y}_{\mathrm{Ie}}$, the Taylor expansion shows that

$$\widehat{Y}_{\mathrm{Ie}} = \widehat{Y}_{\mathrm{Id}} + (\hat{\alpha} - \alpha)^\top \Delta + o_p(n^{-1/2} N), \tag{28}$$

where

$$\Delta = \sum_{i=1}^{N} \pi_i (\partial \pi_i^{-1} / \partial \alpha)(y_i - \gamma_M^* x_i).$$

Thus, if $\partial\pi_i^{-1}/\partial\alpha$ is uniformly bounded, the asymptotic unbiasedness of $\widehat{Y}_{\rm Ie}$ follows from (14) and (26), regardless of whether the ratio imputation model is true or not. When the ratio imputation model is true, the expectation of $\Delta$ term is zero and $(\hat\alpha-\alpha)^\top\Delta$ is of a smaller order than $\widehat{Y}_{\rm Id}-Y$. Thus, $\widehat{Y}_{\rm Ie}$ is asymptotically equivalent to $\widehat{Y}_{\rm Id}$ in this case. If the ratio model is not true, the term $(\hat\alpha-\alpha)^\top\Delta$ is no longer negligible and the asymptotic behavior of $\widehat{Y}_{\rm Ie}$ can be different from that of $\widehat{Y}_{\rm Id}$.

For the variance estimation, let $\hat\alpha^{(k)}$ be the $k$th replicate of $\alpha$ satisfying

$$\sum_{k=1}^{L} c_k(\hat\alpha^{(k)}-\hat\alpha)(\hat\alpha^{(k)}-\hat\alpha)^\top = \mathrm{cov}(\hat\alpha\,|\,\mathcal{F})+o_p(n^{-1}). \tag{29}$$

If we define the $k$th replicate of $\hat\pi_i$ to be $\hat\pi_i^{(k)}=\pi(x_i;\,\hat\alpha^{(k)})$, the proposed variance estimator for $\widehat{Y}_{\rm Ie}$ is

$$\widehat{V}_e=\sum_{k=1}^{L} c_k(\widehat{Y}_{\rm Ie}^{(k)}-\widehat{Y}_{\rm Ie})^2, \tag{30}$$

where

$$\begin{aligned}\widehat{Y}_{\rm Ie}^{(k)} &= \sum_{i\in A} w_i^{(k)}\{R_i y_i+(1-R_i)x_i\hat\gamma_{\rm Me}^{(k)}\}\\ &= \sum_{i\in A} w_i^{(k)}x_i\hat\gamma_{\rm Me}^{(k)}+\sum_{i\in A} w_i^{(k)}(\hat\pi_i^{(k)})^{-1}R_i(Y_i-x_i\hat\gamma_{\rm Me}^{(k)}),\end{aligned}$$

and

$$\hat\gamma_{\rm Me}^{(k)}=\left[\sum_{i\in A} w_i^{(k)}\{(\hat\pi_i^{(k)})^{-1}-1\}R_ix_i\right]^{-1}\sum_{i\in A} w_i^{(k)}\{(\hat\pi_i^{(k)})^{-1}-1\}R_iY_i.$$

For the consistency of $\widehat{V}_e$ in (30), a Taylor expansion can be used to show that

$$c_k^{1/2}(\widehat{Y}_{\rm Ie}^{(k)}-\widehat{Y}_{\rm Ie})=c_k^{1/2}\sum_{i\in A}(w_i^{(k)}-w_i)\zeta_i+c_k^{1/2}(\hat\alpha^{(k)}-\hat\alpha)^\top\Delta+o_p(L^{-1/2}n^{-1/2}N),$$

where $\zeta_i$ is defined following (23) and $\Delta$ is defined following (28). Thus, by (19) and (29), $\widehat{V}_e$ is consistent for the conditional variance of $\widehat{Y}_{\rm Ie}$, conditional on $R_1,\ldots,R_N$. The consistency for the unconditional variance also follows using the same argument for the consistency of $\widehat{V}_d$ in Section 4.

## 6. SIMULATION STUDY

To test our theory, we performed a limited simulation study. The simulation study can be described as a $2\times 3$ factorial design with $B=10{,}000$ replication within each cell. The factors are two types of sampling distribution and three types of imputed estimator. For the sampling distribution, one is generated by a ratio model and the other is generated by a nonratio model. In the ratio model, we generated

$$y_i=3.9x_i+\sqrt{x_i}\,\varepsilon_i,$$

where $x_i\sim\mathcal{U}(0.1,2.1)$, $\varepsilon_i\sim\mathsf{N}(0,1)$, and $x_i$ and $\varepsilon_i$ are independent. In the nonratio model, we used the same $x_i$ and $\varepsilon_i$, but the $y_i$ are generated differently:

$$y_i=(1.8x_i-1)^2+\sqrt{x_i}\,\varepsilon_i.$$

Two sets of random sample of size $n=100$ are separately generated from the two infinite populations. For the response probabilities $\pi_i$, we use the logistic model

$$\pi_i=\exp(-1+2.3x_i)/\{1+\exp(-1+2.3x_i)\}$$

and the overall response rates are all 0.76. The regression coefficients of the logistic model are estimated by the maximum likelihood method and computed iteratively using the Newton–Raphson method.

From each simulated value $(x_i, \pi_i, R_i, y_i)$, $i = 1, \ldots, n$, we computed three types of imputed estimator of the population mean $\theta$ of $y$: $\hat{\theta}_I$ using $y_i^* = x_i \hat{\gamma}_R$ in (4), $\hat{\theta}_{\text{Id}}$ using $y_i^* = x_i \hat{\gamma}_{M2}$ in (7), and $\hat{\theta}_{\text{Ie}}$ using $y_i^* = x_i \hat{\gamma}_{\text{Me}}$ in (27). We also computed the complete sample estimator $\hat{\theta}_n$ for comparison. For the variance estimator, we used the standard jackknife method, where $c_k$ is $n^{-1}(n-1)$ and $w_i^{(k)}$ is defined as $w_i^{(k)} = (n-1)^{-1} n w_i$ if $i \neq k$, and $w_i^{(k)} = 0$ if $i = k$.

Table 1 shows the mean, the variance, and the standardized variance, and the standardized mean squared error (MSE) of the point estimators. The standardized variance and the standardized MSE are the relative variance and the relative MSE compared with those of $\hat{\theta}_n$, respectively. Under the ratio model, all the point estimators are unbiased and the imputed estimator using $\hat{\theta}_I$ is slightly more efficient than $\hat{\theta}_{\text{Id}}$ or $\hat{\theta}_{\text{Ie}}$. In fact, under simple random sampling, the estimator $\hat{\theta}_I$ is the best linear unbiased predictor for the mean of $y$ under the ratio model. However, under the nonratio model, $\hat{\theta}_I$ is biased because the ratio imputation model is no longer true. The estimator $\hat{\theta}_{\text{Ie}}$ is still unbiased and is slightly more efficient than $\hat{\theta}_{\text{Id}}$ because $\hat{\theta}_{\text{Ie}}$ uses additional information that is not captured by the ratio imputation.

TABLE 1: Means, variances, standardized variances, and standardized mean squared errors of the point estimators, based on 10,000 samples.

| Model | Method | Mean | Variance | Standardized Variance | Standardized MSE |
|---|---|---|---|---|---|
| Ratio Model | $\hat{\theta}_n$ | 4.29 | 0.061 | 100 | 100 |
| | $\hat{\theta}_I$ | 4.29 | 0.062 | 102 | 102 |
| | $\hat{\theta}_{\text{Id}}$ | 4.29 | 0.063 | 104 | 104 |
| | $\hat{\theta}_{\text{Ie}}$ | 4.29 | 0.063 | 104 | 104 |
| Nonratio Model | $\hat{\theta}_n$ | 2.04 | 0.062 | 100 | 100 |
| | $\hat{\theta}_I$ | 2.20 | 0.067 | 108 | 148 |
| | $\hat{\theta}_{\text{Id}}$ | 2.04 | 0.067 | 107 | 107 |
| | $\hat{\theta}_{\text{Ie}}$ | 2.04 | 0.066 | 106 | 106 |

Table 2 shows the relative mean and $t$-statistic for the estimated variances of four point estimators. The relative mean of the estimated variance is the Monte Carlo mean of the estimated variance divided by the Monte Carlo variance of the point estimator. The $t$-statistic is the Monte Carlo estimated bias divided by the Monte Carlo standard error of the estimated bias. The simulation result shows that all the variance estimators are asymptotically unbiased under the ratio model. However, under the nonratio model, the estimator $\widehat{V}_I$ shows significant biases when the assumed imputation model does not hold. Both $\widehat{V}_d$ and $\widehat{V}_e$ show nonsignificant biases because they properly take the response model into account.

The above simulation results show that the proposed estimators are unbiased and exhibit good finite-sample performances when the assumed response model is incorrect. We performed another simulation study where the assumed response model is false. In that simulation, we set the true response probability to be $\pi_i = 1 - 0.7(x_i - 1.1)^2$ and used the logistic regression model to estimate the response probability. Thus, the assumed response model is not true in this simulation. The simulation results, although not reported here, show that the estimators $\hat{\theta}_{\text{Ie}}$ and $\widehat{V}_e$ are still unbiased under the ratio model, but are biased under the nonratio model. These results are consistent with the argument of the double protection of the proposed estimator: When either the response model or the imputation model is true, the proposed estimators are unbiased and thus doubly protected against the failure of the assumed models.

TABLE 2: Relative means and $t$-statistics for the variance estimators, based on 10,000 samples.

| Model | Method | Relative Mean | $t$-statistic |
|---|---|---|---|
| Ratio Model | $\widehat{V}_I$ | 1.02 | 1.74 |
| | $\widehat{V}_d$ | 1.02 | 1.55 |
| | $\widehat{V}_e$ | 1.02 | 1.69 |
| Nonratio Model | $\widehat{V}_I$ | 0.73 | -19.93 |
| | $\widehat{V}_d$ | 1.00 | 0.25 |
| | $\widehat{V}_e$ | 1.01 | 0.38 |

## 7. CONCLUDING REMARKS

The concept of *doubly protected imputation* against the failure of the assumed models is relatively a new concept. In some literature, it is called “doubly robust”, as in Scharfstein, Rotnitsky & Robins (1999) and Van der Laan & Robins (2003). The basic motivation is that since one is never sure whether either a response model or a imputation model is correct, perhaps the best that can be hoped for is to find an estimator that can be justified under either one of the two models so that the resulting estimator gives the analyst two chances, instead of only one, to make a valid inference (Bang & Robins 2005).

Here, we have described the same property as doubly protected because the traditional so-called “robust estimator” is not too sensitive to departures from model assumptions in the inferential context chosen. One contribution of this study is to extend the doubly protected inference into the imputation context. However, as noted by an anonymous referee, the proposed imputation method is not doubly protected for domain estimation. Generally speaking, a single set of imputed data ignoring the domain information leads to biased estimation for the domains. Recently, Haziza & Rao (2005) have proposed a bias-correction method for domain estimation after imputation. Domain estimation after imputation, although very important in practice, is beyond the scope of this paper and is not discussed here. So far, we have only considered the ratio imputation. Extensions to other imputation methods, such as regression imputation and hot deck imputation, are not discussed here and will be topics for future research.

## APPENDIX

*Proof of* $(13)$. We first express $\widehat{Y}_{\mathrm{Id}}$ as

$$\widehat{Y}_{\mathrm{Id}} = \widehat{Y}_R + \widehat{X}_M \frac{\widehat{Y}^* - \widehat{Y}_R}{\widehat{X}^* - \widehat{X}_R}, \tag{A.1}$$

where

$$(\widehat{X}^*, \widehat{Y}^*) = \sum_{i \in A} w_i \pi_i^{-1} R_i (x_i, y_i), \quad (\widehat{X}_R, \widehat{Y}_R) = \sum_{i \in A} w_i R_i (x_i, y_i)$$

and

$$\widehat{X}_M = \sum_{i \in A} w_i (1 - R_i) x_i.$$

By (9) and (10),

$$\mathrm{E}\left\{ (\widehat{X}_R - X_R)^2 \,|\, \mathcal{F}, R_1, \ldots, R_N \right\} = O(n^{-1} N^2). \tag{A.2}$$

Thus, using Corollary 5.1.1.2 of Fuller (1996), we have

$$N^{-1}(\widehat{X}_R - X_R) = O_p(n^{-1/2}). \tag{A.3}$$

Similarly, we have

$$N^{-1}(\widehat{Y}_R - Y_R, \widehat{X}_M - X_M) = O_p(n^{-1/2}) \tag{A.4}$$

and

$$N^{-1}(\widehat{X}^* - X^*, \widehat{Y}^* - Y^*) = O_p(n^{-1/2}). \tag{A.5}$$

Hence, by (A.3), (A.4), and (A.5), we can apply the Taylor expansion on (A.1) to get

$$\widehat{Y}_{\mathrm{Id}} = \widehat{Y}_R + \gamma_M^* \widehat{X}_M + \delta\big\{\widehat{Y}^* - \widehat{Y}_R - \gamma_M^*(\widehat{X}^* - \widehat{X}_R)\big\} + o_p(n^{-1/2}N).$$

Therefore, since $\widehat{Y}_n = \widehat{Y}_R + \widehat{Y}_M$, result (13) follows.

*Proof of* (23). Write

$$\widehat{Y}_{\mathrm{Id}}^{(k)} = \widehat{Y}_R^{(k)} + \widehat{X}_M^{(k)} \frac{\widehat{Y}^{*(k)} - \widehat{Y}_R^{(k)}}{\widehat{X}^{*(k)} - \widehat{X}_R^{(k)}}, \tag{A.6}$$

where

$$(\widehat{X}^{*(k)}, \widehat{Y}^{*(k)}) = \sum_{i \in A} w_i^{(k)} \pi_i^{-1} R_i(x_i, y_i), \quad (\widehat{X}_R^{(k)}, \widehat{Y}_R^{(k)}) = \sum_{i \in A} w_i^{(k)} R_i(x_i, y_i)$$

and

$$\widehat{X}_M^{(k)} = \sum_{i \in A} w_i^{(k)} (1 - R_i) x_i.$$

By (10) and (22),

$$N^{-1} c_k^{1/2} (\widehat{X}_n^{(k)} - \widehat{X}_n, \widehat{Y}_n^{(k)} - \widehat{Y}_n) = O_p(n^{-1/2} L^{-1/2}). \tag{A.7}$$

Also, by (A.2) and (22), conditional on $R_1, \ldots, R_N$,

$$N^{-1} c_k^{1/2} \big(\widehat{X}_R^{(k)} - \widehat{X}_R, \widehat{Y}_R^{(k)} - \widehat{Y}_R\big) = O_p(n^{-1/2} L^{-1/2}). \tag{A.8}$$

Similarly, conditional on $R_1, \ldots, R_N$,

$$N^{-1} c_k^{1/2} \big(\widehat{X}^{*(k)} - \widehat{X}^*, \widehat{Y}^{*(k)} - \widehat{Y}^*\big) = O_p(n^{-1/2} L^{-1/2}) \tag{A.9}$$

and

$$N^{-1} c_k^{1/2} (\widehat{X}_M^{(k)} - \widehat{X}_M) = O_p(n^{-1/2} L^{-1/2}). \tag{A.10}$$

By a Taylor expansion on (A.6), using (A.7)–(A.10),

$$\begin{aligned} c_k^{1/2}(\widehat{Y}_{\mathrm{Id}}^{(k)} - \widehat{Y}_{\mathrm{Id}}) \quad = \quad & c_k^{1/2}(\widehat{Y}_R^{(k)} - \widehat{Y}_R^{(k)}) + c_k^{1/2} \hat{\gamma}_M^* (\widehat{X}_M^{(k)} - \widehat{X}_M) \\ & + c_k^{1/2} \hat{\delta} \Big[(\widehat{Y}^{*(k)} - \widehat{Y}^*) - (\widehat{Y}_R^{(k)} - \widehat{Y}_R) \\ & - \hat{\gamma}_M^* \big\{(\widehat{X}^{*(k)} - \widehat{X}^*) - (\widehat{X}_R^{(k)} - \widehat{X}_R)\big\}\Big] \\ & + o_p(L^{-1/2} n^{-1/2} N), \end{aligned} \tag{A.11}$$

where $\hat{\gamma}_M^* = (\widehat{X}^* - \widehat{X}_R)^{-1}(\widehat{Y}^* - \widehat{Y}_R)$ and $\hat{\delta} = (\widehat{X}^* - \widehat{X}_R)^{-1} \widehat{X}_M$. Using (A.3), (A.4), and (A.5), a Taylor expansion can be used to show that

$$\hat{\gamma}_M^* = \gamma_M^* + o_p(n^{-1/2}) \tag{A.12}$$

and

$$\hat{\delta} = \delta + o_p(n^{-1/2}). \tag{A.13}$$

Therefore, result (23) follows by inserting (A.12) and (A.13) into (A.11).

## ACKNOWLEDGEMENTS

The work of the first author was supported by the Korea Research Foundation Grant (KRF–2004–041–C00072). The authors thank an anonymous referee for helpful comments on the original version of the manuscript.

## REFERENCES

H. Bang & J. M. Robins (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61, 962–973.

K. R. W. Brewer (1979). A class of robust sampling designs for large-scale surveys. *Journal of the American Statistical Association*, 74, 911–915.

R. E. Fay (1991). A design-based perspective on missing data variance. In *Bureau of the Census: 1991 Annual Research Conference*, U.S. Bureau of the Census, Washington, DC, pp. 429–440.

W. A. Fuller (1996). *Introduction to Statistical Time Series*, Second Edition. Wiley, New York.

R. Groves, D. Dillman, J. Eltinge & R. J. A. Little (2002). *Survey Nonresponse*. Wiley, New York.

D. Haziza & J. N. K. Rao (2005). Inference for domains under imputation for missing survey data. *The Canadian Journal of Statistics*, 33, 149–161.

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

G. Kalton (1983). *Compensating for Missing Survey Data*. Institute for Social Research, University of Michigan, Ann Arbor, MI.

J. K. Kim, A. Navarro & W. A. Fuller (2006). Replication variance estimation for two-phase stratified sampling. *Journal of the American Statistical Association*, in press.

S. R. Lipsitz, J. G. Ibrahim & L. P. Zhao (1999). A weighted estimating equation for missing covariate data with properties similar to maximum likelihood. *Journal of the American Statistical Association*, 94, 1147–1160.

J. N. K. Rao (1996). On variance estimation with imputed survey data. *Journal of the American Statistical Association*, 91, 499–506.

J. N. K. Rao & J. Shao (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79, 811–822.

J. N. K. Rao & R. R. Sitter (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, 82, 453–460.

J. M. Robins, A. Rotnitzky & L. P. Zhao (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89, 846–866.

P. R. Rosenbaum (1987). Model-based direct adjustment. *Journal of the American Statistical Association*, 82, 387–394.

C. E. Särndal, B. Swensson & J. Wretman (1992). *Model Assisted Survey Sampling*. Springer, New York.

D. Scharfstein, A. Rotnitsky & J. Robins (1999). Adjusting for nonignorable dropout using semiparametric models (with discussion). *Journal of the American Statistical Association*, 94, 1096–1146.

J. Shao & P. Steel (1999). Variance estimation for survey data with composite imputation and non-negligible sampling fractions. *Journal of the American Statistical Association*, 94, 254–265.

M. J. Van der Laan & J. M. Robins (2003). *Unified Methods for Censored Longitudinal Data and Causality.* Springer, New York.

*Received 3 February 2005*
*Accepted 4 August 2005*

Jae Kwang KIM: kimj@yonsei.ac.kr
*Department of Applied Statistics*
*Yonsei University*
*Seoul, Korea 120–749*

Hyeonah PARK: chang@stats.snu.ac.kr
*Department of Statistics*
*Seoul National University*
*Seoul, Korea 151–742*