Fractional imputation using regression imputation model

Jae Kwang Kim
Department of Statistics, Hankuk University of FS, Yongin, 449-791, KOREA

Key Words: Jackknife, Nonresponse, Response probability, Survey sampling, Variance Estimation.

1. Introduction

Consider a finite population of N elements identified by a set of indices $U = \{1, 2, ..., N\}$. Associated with each unit i in the population there is a study variable y_i and a vector $\mathbf{x}_i$ of auxiliary variables. Let A denote the set of indices for the elements in a sample selected by a set of probability rules called the *sampling mechanism*. Let the population quantity of interest be $\theta_N = \sum_{i=1}^N y_i$ or $\theta_N = N^{-1}\sum_{i=1}^N y_i$ and let $\hat{\theta}_n$ be a linear estimator of θ_N based on the full sample,

$$\hat{\theta}_n = \sum_{i \in A} w_i y_i \tag{1}$$

To deal with item nonresponse, we define A_R and A_M as the set of indices of the sample respondents and sample nonrespondents, respectively. Let $R_i = 1$ if unit i belongs to A_R and $R_i = 0$ if unit i belongs to A_M. For each unit with a missing value, we impute to complete the data set and denote the imputed value as y_i^*. Let $\hat{\theta}_I$ denote the imputed estimator, so the imputed estimator of the mean, using the form of the full sample estimator given by (1), is

$$\hat{\theta}_I = \sum_{i \in A_R} w_i y_i + \sum_{i \in A_M} w_i y_i^*. \tag{2}$$

In many practical cases, the imputed value y_i^* is written as a predicted value plus a residual term

$$y_i^* = \hat{y}_i + \hat{e}_i^*, \tag{3}$$

where $\hat{y}_i$ is the predicted value of y_i and $\hat{e}_i^*$ is an imputed residual selected at random from $\{\hat{e}_i = y_i - \hat{y}_i;\ i \in A_R\}$ in the same cell. When the predicted value for unit i is $\hat{y}_i = \mathbf{x}_i'\hat{\boldsymbol{\beta}}$ with $\hat{\boldsymbol{\beta}} = \left(\sum_{i \in A_R} w_i \mathbf{x}_i \mathbf{x}_i'\right)^{-1} \sum_{i \in A_R} w_i \mathbf{x}_i y_i$, we call the imputation method defined in (3) *stochastic regression imputation*. The representation in (3) is a general form and it covers many commonly used imputation procedures such as hot deck imputation or ratio imputation. (Rao, 1996).

For variance estimation of the imputed estimator, the adjusted jackknife method is often used. Rao and Shao (1992) introduced the method, applying it for a weighted hot-deck where the donors are selected with-replacement with the selection probability proportional to their weights. Rao (1996) discussed the adjusted jackknife method in detail for various imputation methods, but did not cover stochastic regression imputation.

Rao and Shao (1992) and Fuller and Kim (2002) studied asymptotic properties of the random hot deck imputation method in detail under the response probability model. The response probability model does not require a correct specification of underlying population distribution and is often preferred to other model-based approaches (Fay, 1996). In this paper, we extend Fuller and Kim (2002) result to stochastic regression imputation. A new variance estimator, slightly different from the adjusted jackknife method, is introduced and its asymptotic properties are discussed. Shao and Steel (1999) also covered complex imputation methods including the stochastic regression imputation but, as we shall see in Section 4, their variance estimator shows limited performance compared to the new variance estimator proposed in this article.

The paper is organized as follows. Asymptotic properties of the imputed estimator is derived under the uniform response model in Section 2. In Section 3, a new variance estimation method is discussed under the uniform response model. In Section 4, concluding remarks are made with some simulation results.

2. PROPERTIES OF IMPUTED POINT ESTIMATOR

Rao and Shao (1992) proposed the adjusted jackknife method and examine its statistical properties under the response probability model. Under this model, response is treated as a second phase of sampling from the complete sample and inferences are under the joint distribution of the sampling distribution, the assumed response mechanism, and the imputation mechanism. The response model they use assumes the probability of response is constant within each cell used for imputing. They show that if the missing values are imputed with a weighted hot deck then the imputed estimator is asymptot-

ically unbiased and the adjusted jackknife method gives an asymptotically unbiased estimator of the variance of the imputed estimator. Their development applies for both simple random samples and stratified multistage sampling with ignorable sample rate.

A key idea that Rao and Shao (1992) used in their development of the adjusted jackknife can be appreciated by decomposing the imputed estimator. We can write the imputed estimator as

$$\hat{\theta}_I = \hat{\theta}_n + \left\{ E_I\left(\hat{\theta}_I\right) - \hat{\theta}_n \right\} + \left\{ \hat{\theta}_I - E_I\left(\hat{\theta}_I\right) \right\}, \quad (4)$$

where the expectation $E_I(\cdot)$ denotes the expectation over the imputation mechanism. The variance of the imputed estimator is then given by

$$\begin{aligned} Var\left(\hat{\theta}_I\right) &= Var\left(\hat{\theta}_n\right) + Var\left(E_I\left(\hat{\theta}_I\right) - \hat{\theta}_n\right) \\ &\quad + Var\left(\hat{\theta}_I - E_I\left(\hat{\theta}_I\right)\right) \end{aligned} \quad (5)$$

provided all the covariance terms are zero. In expression (5), the variances are over the sample design, the response mechanism, and the imputation mechanism. The first term on the right hand side of (5) is the sampling variance, the second term is the variance due to response machanism, and the third term is the imputation variance due to the selection of a random donor.

Two of the three covariances arising from (4) are easily shown to be equal to zero because the third term $\hat{\theta}_I - E_I\left(\hat{\theta}_I\right)$ has zero expectation over the imputation mechanism. Only the term $Cov\left(\hat{\theta}_n, E_I\left(\hat{\theta}_I\right) - \hat{\theta}_n\right)$ requires additional investigation. Rao and Shao (1992) show that this covariance is asymptotically equal to zero if two conditions are met when a hot deck is used to impute for missing values. One condition is that the donors for the hot deck are imputed with probabilities proportional to their weights. The second condition is that the sampled units have the same probability of responding within cells. We extend the results given by Rao and Shao (1992) to a slightly more general setting of stochastic regression imputation.

Let d_{ij} be an indicator function that takes the value one if unit i is used as a donor for missing unit j. The distribution of the d_{ij} is called the imputation mechanism. The imputed estimator of the population total given by (2) can be written as

$$\hat{\theta}_I = \sum_{i \in A_R} w_i y_i + \sum_{j \in A_M} w_j \left(\hat{y}_j + \sum_{i \in A_R} d_{ij} \hat{e}_i \right). \quad (6)$$

An imputed estimator is asymptotically, conditionally unbiased if

$$E_R E_I\left(\hat{\theta}_I\right) - \hat{\theta}_n = o_p\left(n^{-1/2}\right). \quad (7)$$

Under the within-cell uniform response model, the following lemma gives a necessary and sufficient condition for an imputed estimator to be asymptotically unbiased in the sense given by (7). This extends the results of Rao and Shao (1992) by showing the condition is also necessary. It also broadens the imputation procedures covered to include regression imputation, as defined by (6).

Lemma 2..1 *Let the complete sample estimator be of the form (1). Assume the imputed estimator is a member of the linear class defined in (6). Under the within-cell uniform response model, a necessary and sufficient condition for an imputed estimator to satisfy (7) is*

$$E_I(d_{ij}) = \begin{cases} \left(\sum_{i \in A_{Rg}} w_i\right)^{-1} w_i & \text{if } i \in A_{Rg} \\ & \text{and } j \in A_{Mg} \\ 0 & \text{otherwise,} \end{cases} \quad (8)$$

where $A_{Rg} = A_R \cap U_g$, $A_{Mg} = A_M \cap U_g$, and the expectation is taken with respect to the imputation mechanism.

Proof. Under the uniform within-cell response model, the response probability is constant within a cell. We denote this probability as $\pi_g = Pr(i \in A_{Rg} \mid i \in A_g)$. From (6),

$$\begin{aligned} & E_R E_I\left(\hat{\theta}_I - \hat{\theta}_n\right) \\ = \; & E_R E_I\left(\sum_{i \in A} \sum_{k \in A} w_i (1 - R_i) R_k d_{ki} \hat{e}_k - \sum_{i \in A} w_i (1 - R_i) \hat{e}_i \right) \\ = \; & \sum_{g=1}^{G} (1 - \pi_g) \sum_{k \in A_g} E_R(\hat{e}_k) \left(\sum_{i \in A_g} w_i \pi_g E_I(d_{ki}) - w_k \right) \end{aligned}$$

Hence, for the quantity to be equal to zero for any y-variable we need $\sum_{i \in A_g} w_i \pi_g E_I(d_{ki}) - w_k = 0$. This implies that $E_I(d_{ij})$ must be proportional to w_i for i and j in the same cell and $E_I(d_{ij})$ must equal zero if i and j are in different cells. This prove that (8) is a necessary condition. The sufficiency part is proved in Theorem 2..1. ∎

The lemma indicates that if the response probability model is the justification for the imputation procedure, then an unweighted selection method cannot give an unbiased point estimator (except in the equally weighted case). The result applies to both

the within-cell weighted hot deck and the more general (weighted) stochastic regression imputation.

When the imputed estimator is the sum of a predicted value and a residual imputed by a weighted selection (8), then the expected value of any unbiased imputed estimator over the imputation mechanism is

$$E_I\left(\hat{\theta}_I\right) = \sum_{i\in A} w_i\hat{y}_i + \sum_{g=1}^{G}\sum_{i\in A_{Rg}} w_i\hat{\pi}_g^{-1}\hat{e}_i =: \hat{\theta}_{FE} \quad (9)$$

where $\hat{y}_i$ is the predicted value of y_i defined after (3), $\hat{\pi}_g = \left(\sum_{i\in A_g} w_i\right)^{-1}\sum_{i\in A_{Rg}} w_i$ is the estimated response probability of group g, and the notation $A =: B$ means that B is defined to be equal to A. The hypothetical estimator $\hat{\theta}_{FE}$ is approximately unbiased for the population mean of y_i's and does not involves the extra variance due to random imputation. The subscript FE is the abbreviation of "fully efficient" in the sense that it has the smallest variance among the imputed estimator satisfying (8).

The estimator (9) can be implemented by using fractional imputation in which every responding unit in an imputation cell is used as a donor for every nonrespondent in the cell, and the imputation weight is proportional to the sampling weight. Fractional imputation is first proposed by Kish and Kalton (1984) and later investigated by Kim and Fuller (2003). Then, the estimator (9) can be written as the fractionally imputed estimator

$$\hat{\theta_{FEFI}} = \sum_{g=1}^{G}\sum_{j\in A_g} w_j\left(\hat{y}_j + \sum_{i\in A_{Rg}} w_{ij}^*\hat{e}_i\right) \quad (10)$$

where $w_j w_{ij}^*$ is the weight of donor i for recipient j, w_{ij}^* is the imputation fraction of donor i for recipient j with

$$w_{ij}^* = \begin{cases} \left(\sum_{s\in A_{Rg}} w_s\right)^{-1} w_i R_i & \text{if } R_j = 0 \\ 1 & \text{if } R_j = 1 \text{ and } i = j. \end{cases} \quad (11)$$

The estimator (9) with w_{ij}^* of (11), algebraically equivalent to (9), is called the *fully efficient fractionally imputed* (FEFI) estimator.

In the following theorem, we establish the properties of the regression imputation estimator given by (9) under the within-cell uniform response model. Appendix contains a more complete statement and the proof of this theorem.

Theorem 2..1 *Assume the same structure of the estimator and the population as in Lemma 2.1 and some regularity conditions as explicitly stated in Appendix B. Assume for every $i \neq j = 1, 2, \cdots, N$,*

$$P(R_i = 1, R_j = 1) = P(R_i = 1)P(R_j = 1), \quad (12)$$

where R_i is the response indicator function of unit i. Let the predictor $\hat{y}_i$ of unit i be of the form $\hat{y}_i = \mathbf{x}_i'\hat{\boldsymbol{\beta}}$ where $\hat{\boldsymbol{\beta}}$ satisfies

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = O_p\left(n^{-1/2}\right) \quad (13)$$

for some population value $\boldsymbol{\beta}_0$. Then,

$$\hat{\theta}_{FE} = \hat{\theta}_n + \sum_{g=1}^{G}\sum_{i\in A_g}\left(\pi_g^{-1}R_i - 1\right) w_i e_{ig} + o_p(n^{-1/2}), \quad (14)$$

where $\hat{\theta}_{FE}$ is an estimator of population mean defined by (9), $\pi_g = Pr\left(R_i = 1 \mid i \in A_g\right)$ is the response probability in cell g, $e_{ig} = y_i - \bar{Y}_g - \left(\mathbf{x}_i' - \bar{\mathbf{X}}_g{}'\right)\boldsymbol{\beta}_0$, and $\left(\bar{\mathbf{X}}_g, \bar{Y}_g\right)$ is the population mean of $(\mathbf{x}_i, y_i)$ in cell g.

For the proof, see Appendix.

Theorem 2..1 shows that under the uniform within-cell response model, the second term on the right hand side of (14) has zero expectation. As a result, it follows that

$$E_R\left(\hat{\theta}_{FE}\right) \doteq \hat{\theta}_n. \quad (15)$$

The approximate equality given by (15) implies the covariance term arising from decomposition (4) is

$$Cov\left(\hat{\theta}_n, \hat{\theta}_{FE} - \hat{\theta}_n\right) \doteq 0.$$

Since all the covariance terms are equal to zero provided we have a uniform, within-cell response model and the donors are selected with probability proportional to their weights, the variance given by (5) is valid. The first component of variance in the expression is the ordinary sampling variance of the complete sample. The second term, the variance due to response mechanism, is can be further developed. Using expression (9) and the conditional unbiasedness given by (15), this variance is

$$Var\left(\hat{\theta}_{FE} - \hat{\theta}_n\right) \doteq E_p\left[Var_R\left\{\sum_{g=1}^{G}\sum_{i\in A_g}\left(\pi_g^{-1}R_i - 1\right) w_i e_{ig}\right\}\right]. \quad (16)$$

By the independence assumption of (12) and the fact that the variance of R_i is $\pi_g\left(1 - \pi_g\right)$, the variance

reduces further to

$$Var\left(\hat{\theta}_{FE}-\hat{\theta}_n\right) \doteq E_p\left\{\sum_{g=1}^{G}\pi_g^{-1}\left(1-\pi_g\right)\sum_{i\in A_g}w_i^2e_{ig}^2\right\}. \quad (17)$$

Expression (17) makes it clear that the response probabilities and the weighted residuals are the key factors that determine the variance due to response mechanism. If the π_g are all close to unity, then the variance will be close to zero. On the other hand, response probabilities close to zero will make the variance due to response mechanism large. The residuals also affect the variance. If the imputation model is good in the sense that the predicted values are close to the actual values, then the sum of the squared weighted residuals and the variance due to response mechanism will be small.

3. VARIANCE ESTIMATION

We now consider variance estimation. Under complete response, let a replication variance estimator be

$$\hat{V}(\hat{\theta}_n) = \sum_{k=1}^{L}c_k(\hat{\theta}_n^{(k)}-\hat{\theta}_n)^2, \quad (18)$$

where $\hat{\theta}_n^{(k)}$ is the k-th estimate of θ_N 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 $\hat{\theta}_n$ is a linear estimator of the form (1), the k-th replicate of $\hat{\theta}_n$ can be written

$$\hat{\theta}_n^{(k)} = \sum_{i\in A}w_i^{(k)}y_i \quad (19)$$

where $w_i^{(k)}$ denotes the replicate weight for the i-th unit of the k-th replication. For example, consider a simple random sample of size n with $w_i\equiv n^{-1}$. Then, a replication variance estimator is the jackknife variance estimator defined by $L=n$, $c_k=n^{-1}(n-1)$ and $w_i^{(k)}=(n-1)^{-1}$ if $i\neq k$ and $w_i^{(i)}=0$.

For the stochastic imputation estimator satisfying (8), recall that the total variance can be written as

$$Var\left(\hat{\theta}_I\right)=Var\left(\hat{\theta}_{FE}\right)+Var\left(\hat{\theta}_I-\hat{\theta}_{FE}\right). \quad (20)$$

The two terms can be estimated separately. The first term, the variance over the sampling mechanism and the response mechanism, can be estimated by

$$\hat{V}_{FE}=\sum_{k=1}^{L}c_k\left(\hat{\theta}_{FE}^{(k)}-\hat{\theta}_{FE}\right)^2, \quad (21)$$

where

$$\hat{\theta}_{FE}^{(k)}=\sum_{i\in A}w_i^{(k)}\hat{y}_i+\sum_{g=1}^{G}\sum_{i\in A_{Rg}}w_i^{(k)}\left[\hat{\pi}_g^{(k)}\right]^{-1}\hat{e}_i, \quad (22)$$

and $\hat{\pi}_g^{(k)}=\left(\sum_{i\in A_g}w_i^{(k)}\right)^{-1}\sum_{i\in A_{Rg}}w_i^{(k)}$.

Under some conditions, it can be shown that the replicates defined in (22) satisfies

$$\begin{aligned}\hat{\theta}_{FE}^{(k)}-\hat{\theta}_{FE} &= \hat{\theta}_n^{(k)}-\hat{\theta}_n \quad (23)\\ &+\sum_{g=1}^{G}\sum_{i\in A_g}\left(w_i^{(k)}-w_i\right)\left(\frac{R_i}{\pi_g}-1\right)e_{ig}\\ &+o_p\left(n^{-1}\right),\end{aligned}$$

where $e_{ig}=y_i-\bar{Y}_g-\left(\mathbf{x}_i'-\bar{\mathbf{X}}_g'\right)\boldsymbol{\beta}_0$ and $\left(\bar{\mathbf{X}}_g,\bar{Y}_g\right)$ is the population mean of $(\mathbf{x}_i,y_i)$ in cell g.

Using (23) in conjunction with (14), we can re-express the replicate variance estimator (for $\hat{\theta}_{FE}$) as

$$\begin{aligned}&\sum_{k=1}^{L}c_k\left(\hat{\theta}_{FE}^{(k)}-\hat{\theta}_{FE}\right)^2\doteq\sum_{k=1}^{L}c_k\left(\hat{\theta}_n^{(k)}-\hat{\theta}_n\right)^2 \quad (24)\\ &+\sum_{k=1}^{L}c_k\left[\sum_{g=1}^{G}\sum_{i\in A_g}\left(w_i^{(k)}-w_i\right)\left(\frac{R_i}{\pi_g}-1\right)e_{ig}\right]^2\\ &+(\text{Cross Product}).\end{aligned}$$

The first term on the right hand side estimates the sampling variance, the second term estimates the conditional variance of $\sum_{g=1}^{G}\sum_{i\in A_g}w_i\left(\pi_g^{-1}R_i-1\right)e_{ig}$, conditional on the values of R_i in the population. By (14), the second term estimates the conditional variance of $\hat{\theta}_{FE}-\hat{\theta}_n$. The cross product term in (24) has zero expectation under the uniform within-cell response model. The conditional mean $E\left(\hat{\theta}_{FE}-\hat{\theta}_n\mid R_1,R_2,\cdots,R_N\right)$ has variance of order N^{-1}. Shao and Steel (1999) and Fuller and Kim (2002) discussed the estimation of the variance for the conditional mean $E\left(\hat{\theta}_{FE}-\hat{\theta}_n\mid R_1,R_2,\cdots,R_N\right)$. If the sampling rate is ignorable, then the variance term can be safely ignored.

It is of interest to compare the proposed variance estimator to the variance estimator proposed by Shao and Steel (1999). Under ignorable sampling rate, Shao and Steel (1999) suggest linearizing the imputed estimator treating all the R_i's as fixed so that it can be written as $\sum_{i\in A}w_i\xi_i$ for some ξ_i and then applying the standard variance estimator to the linearized form with ξ_i substituted by $\hat{\xi}_i$ computed from the respondents. The linearization step

is essentially the same as Theorem 2..1;

$$\hat{\theta}_{FE} \doteq \sum_{i \in A} w_i \xi_i$$

where $\xi_i = \mathbf{x}_i' \boldsymbol{\beta}_0 + \bar{e}_g + \pi_g^{-1} R_i \left(e_i - \bar{e}_g\right)$ for unit i in cell g. However, the variance estimation step is different. Shao and Steel (1999) suggested using a substitution $\hat{\xi}_i$ of ξ_i in applying the standard variance formula. A substitution estimator of ξ_i is

$$\begin{aligned} \hat{\xi}_i &= \bar{y}_{Rg} + \left(\mathbf{x}_i - \bar{\mathbf{x}}_{Rg}\right)' \hat{\boldsymbol{\beta}} \\ &+ \hat{\pi}_g^{-1} R_i \left(y_i - \bar{y}_{Rg} - \left(\mathbf{x}_i - \bar{\mathbf{x}}_{Rg}\right)' \hat{\boldsymbol{\beta}}\right), \end{aligned}$$

where

$$\left(\bar{y}_{Rg}, \bar{\mathbf{x}}_{Rg}\right) = \left(\sum_{i \in A_{Rg}} w_i\right)^{-1} \sum_{i \in A_{Rg}} w_i \left(y_i, \mathbf{x}_i\right)$$

and $\hat{\pi}_g = \left(\sum_{i \in A_g} w_i\right)^{-1} \sum_{i \in A_{Rg}} w_i$ is the estimated response probability for cell g. The Shao-Steel variance estimator for $\hat{\theta}_{FE}$ can be obtained by applying the complete sample variance estimator to $\hat{\xi}_i$ in the sample.

Now we consider the estimation of the second term in (20), the imputation variance. Conceptually, the imputation variance can be unbiasedly estimated by $\left(\hat{\theta}_I - \hat{\theta}_{FE}\right)^2$ because the conditional expectation of $\hat{\theta}_I - \hat{\theta}_{FE}$ is equal to zero. However, the degrees of freedom for estimating the imputation variance will then be only one. One might increase the degrees of freedom by repeatedly applying the given imputation mechanism independently $M > 1$ times. This is essentially the idea of multiple imputation, where the imputation variance is estimated by $M^{-1} B_M = M^{-1} \left(M - 1\right)^{-1} \sum_{t=1}^{M} \left(\hat{\theta}_{I,t} - \bar{\theta}_{I,M}\right)^2$ when $\bar{\theta}_{I,M} = M^{-1} \sum_{t=1}^{M} \hat{\theta}_{I,t}$ is the point estimator in use and $\hat{\theta}_{I,t}$ is the imputed estimator based on the t-th repeated application of the given imputation mechanism. Thus, for single imputation $\hat{\theta}_I$, the variance could be estimated by $\hat{V}_{FE} + B_M$.

Another way of increasing the degrees of freedom, instead of repeating the stochastic imputation, is also possible if we estimate the imputation variance separately within a cell. That is, the imputation variance can be estimated by

$$\hat{V}_{imp} = \sum_{g=1}^{G} \left(\hat{\theta}_{Ig} - \hat{\theta}_{FEg}\right)^2, \tag{25}$$

where $\hat{\theta}_{Ig}$ and $\hat{\theta}_{FEg}$ are the portion of $\hat{\theta}_I$ and $\hat{\theta}_{FE}$, respectively, that belong to cell g. Since $\hat{\theta}_{Ig}$ is conditionally unbiased for $\hat{\theta}_{FEg}$, where the conditional expectation is over the imputation mechanism, the variance estimator in (25) is unbiased for the imputation variance for all stochastic regression estimator satisfying (8). The variance estimator (25) is valid whether the imputation mechanism is with-replacement selection or without-replacement selection. Thus, we do not need to specify full imputation mechanism to compute $\hat{V}_{imp}$ in (25).

If the number of cell G is not large enough, then $\hat{V}_{imp}$ in (25) can be unstable and we may need to calculate the imputation variance explicitly to estimate the variance. Note that the imputed estimator satisfies

$$\hat{\theta}_I - \hat{\theta}_{FE} = \sum_{g=1}^{G} \sum_{i \in A_{Rg}} \sum_{j \in A_{Mg}} w_j \left\{d_{ij} - E_I\left(d_{ij}\right)\right\} \hat{e}_i,$$

where $E_I\left(d_{ij}\right)$ is defined in (8). Thus, for example, if the imputation mechanism is with-replacement selection, then

$$V\left(\hat{\theta}_I - \hat{\theta}_{FE}\right) = E\left[\sum_{g=1}^{G} \sum_{i \in A_{Rg}} \sum_{j \in A_{Mg}} w_j^2 \left\{d_{ij} - E_I\left(d_{ij}\right)\right\}^2 \hat{e}_i^2\right]$$

and the imputation variance can be estimated in a straightforward manner.

4. Simulation Study

The main advantage of the estimators justified under the response model approach is that we do not have to make a correct specification of the distribution of y in the sample. To illustrate this, consider a model

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 \left(x_i^2 - 1\right) + e_i \tag{26}$$

where $x_i \overset{i.i.d.}{\sim} N\left(0, 1\right)$, $e_i \overset{i.i.d.}{\sim} N\left(0, 0.16\right)$, and x_i and e_i are independent. We set $\beta_0 = 0$, $\beta_1 = 0.5$ and $\beta_2 = 0.3$. The variable x_i is always observed but the probability that y_i responds is 0.7. A random sample of size $n = 100$ is generated.

For imputation, suppose that we adopted a linear regression model as the imputation model and computed the predictor of unit i by $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, where $\hat{\beta}_i$, $i = 1, 2$, are the ordinary least squares estimates from the simple regression of y on x. The residuals are randomly drawn by a with-replacement sampling within each cell. The cells are formed using the x values. For comparison, we used several values of cell numbers from $G = 1$ to $G = 15$.

The mean and variance of the point estimators and the relative bias of the estimators of variance are calculated. The point estimators of the population mean are all unbiased and are not listed here.

Table 1 presents the variance and the standardized variance of the point estimators and the relative bias of the variance estimators for each cases. The relative bias of $\hat{V}$ as an estimator of the variance of $\bar{y}_I$ is calculated as $\left[Var_B\left(\bar{y}_I\right)\right]^{-1}\left[E_B\left(\hat{V}\right)-Var_B\left(\bar{y}_I\right)\right]$, where the subscript B denotes the distribution generated by the Monte Carlo simulation. In addition to the variance estimator proposed in this article, we also computed the Shao-Steel variance estimator using the substitution method.

Table 1: Variance of the point estimators and Relative Bias (%) of the variance estimators (10,000 samples).

Number of Cells	Variance	Relative Bias (%)	
		New method	Shao-Steel
1	0.008575	1.73	0.35
3	0.007981	3.14	5.78
5	0.007693	2.88	7.70
7	0.007528	3.17	9.95
9	0.007446	3.64	11.60
11	0.007310	5.00	13.03
13	0.007206	6.33	13.81
15	0.007269	5.37	14.18

The following conclusions can be drawn from the simulation

1. All the point estimators are unbiased because the assumption of within-cell equal response probability holds for all cases.

2. There are differences in the efficiency of the point estimators. The variance of the point estimator is larger for smaller number of imputation cells. This is consistent with our theory in the sense that the variance term in (17) will be solely determined by e_{ig}^2 in our simulation setup. Note that the e_{ig} can be written as an cell-mean adjusted form $e_i - \bar{e}_g$, the error from the imputation model minus its cell mean. Thus, the improvement will be substantial if the magnitude of the original error is large and less variable within each cell.

3. The reduction of the variance is not a linear function of G. In Table 1, the reduction is not substantial for large G. This is because there is a lower bound on the variance of imputed estimator. Note that, in the simulation setup, the variance of y is decomposed as

$$\begin{aligned} Var\left(Y_i\right) &= Var\left(\beta_0+\beta_1 x_i\right)+Var\left[\beta_2\left(x_i^2-1\right)\right] \\ &\quad +Var\left(e_i\right). \end{aligned}$$

The first term in the right side of the above equation is the variance of the systematic part we can catch from the imputation model. The third term is the variance of the pure error term that we can never catch even if the imputation model is equal to the true model. The second term, the variance of $\beta_2\left(x_i^2-1\right)$, represents the magnitude of difference between the imputation model variance and the true model variance. When $G=1$, the second term contributes to the imputation model variance so that, conditional on the number of respondents,

$$\begin{aligned} & Var\left(\hat{\theta}_I \mid r\right) \\ = \;& n^{-1}\beta_1^2+\left(2\beta_1^2+0.16\right)\left[r^{-1}+n^{-2}\left(n-r\right)\right] \\ = \;& 0.00838. \end{aligned}$$

If we use multiple imputation cells suitably, the contribution of the imputation model variance will be reduced. The lower bound of the imputed estimator among all class of linear unbiased estimator is

$$\begin{aligned} \inf Var\left(\hat{\theta}_I \mid r\right) &= n^{-1}\left(\beta_1^2+2\beta_1^2\right)+r^{-1}0.16 \\ &= 0.00659, \end{aligned}$$

where the infimum was taken over all unbiased linear estimator of the mean of y.

4. There is a slight bias of the variance estimator. The bias is essentially a type of ratio bias and will be negligible for large sample size. In fact, we actually increased the sample size to $n=200$ and found that all the relative biases are within 1% in absolute values. However, for a moderate sample size, when there are many imputation cells relative to the number of respondents, the bias for the variance estimators will not be negligible, as is the case with $G>10$ and $n=100$ in Table 1.

5. The proposed variance estimator shows better performance than the Shao-Steel variance estimator for multiple cells. Because the proposed variance estimator uses replicated version of $\hat{\pi}_g$ in the variance estimation, the variability of $\hat{\pi}_g$ for $G>1$ is fully captured in the proposed variance estimator.

Imputation cell is often justified as a method of reducing the nonresponse bias of the imputed estimator. This simulation shows that a suitable choice of imputation cells also make the point estimator efficient. Generally speaking, if the number of imputation cells are large, the point estimator is more efficient but the variance estimator will be more biased. Thus, there is a trade-off between the efficiency of the point estimation and the accuracy of the variance estimation in choosing a suitable number of imputation cells. Eltinge and Yansaneh (1997) also discuss the issue. The choice of optimal number of imputation cell is not discussed here and will be a topic of future research.

Appendix

B. Assumptions and proof of Theorem 2..1

Assumptions

Assume a sequence of finite populations as described in Isaki and Fuller (1981). Define $(\mathbf{x}_i, y_i)$ to be a vector of auxiliary variables and an outcome variable defined on the full population and assume each of these variables have bounded fourth moments. Assume the population consists of G mutually exclusive and exhaustive cells, where N_g is the population size, n_g is the sample size, and r_g is the number of respondents in cell g. Assume

$$K_1 G^{-1} < N^{-1} N_g < K_2 G^{-1} \quad \text{for all } g, \tag{B.1}$$

$$G < K_3 n^{\lambda}, \tag{B.2}$$

$$K_4 \leq n w_i \leq K_5, \tag{B.3}$$

$$K_6 \leq \pi_g \quad \text{for all } g \tag{B.4}$$

where $K_1, K_2, \cdots, K_6$ are fixed positive constants, $0 \leq \lambda < 0.5$, π_g is the common response probability in cell g, and the w_i is the weight of unit i. Assume the complete sample estimator $\hat{\theta}_n$ is unbiased for the finite population total and satisfies

$$Var\left(\hat{\theta}_n\right) < K_M Var\left(\hat{\theta}_{SRS,n}\right) \tag{B.5}$$

for a fixed K_M for any y with bounded moments and $\hat{\theta}_{SRS,n}$ is the estimator of θ based on a simple random sample of size n.

Proof

The difference between estimator (9) and the full sample estimator is

$$\begin{aligned}\hat{\theta}_{FE} - \hat{\theta}_n &= \sum_{g=1}^{G} \left(\sum_{i \in A_g} w_i\right)^{-1} \left(\sum_{i \in A_g} w_i \pi_g^{-1} R_i\right) \\ &\times \left\{\left(\sum_{i \in A_g} w_i \pi_g^{-1} R_i \hat{e}_i\right) - \left(\sum_{i \in A_g} w_i \hat{e}_i\right)\right\},\end{aligned}$$

where R_i is the response indicator function of unit i, π_g is the unknown response probability in cell g, and $\hat{e}_i = y_i - \hat{y}_i$. In order to work with means, we let w_i be the inverse of the initial selection probability divided by N. Now

$$E\left\{\sum_{i \in A_g} w_i\right\} = E\left\{\sum_{i \in A_g} \pi_g^{-1} R_i w_i\right\} = N^{-1} N_g,$$

$$\left[E\left\{\sum_{i \in A} w_i\right\}\right]^{-1} E\left\{\sum_{i \in A_g} w_i\right\} = N^{-1} N_g =: \bar{z}_{Ng},$$

$$E\left\{\sum_{i \in A_g} \pi_g^{-1} w_i R_i e_{ig}\right\} = N^{-1} \sum_{i \in U_g} e_{ig} =: \bar{e}_{Ng},$$

$$E\left\{\sum_{i \in A_g} w_i e_{ig}\right\} = \bar{e}_{Ng},$$

where $\bar{z}_{Ng}$ is the fraction of the population in cell g and $e_{ig} = y_i - \bar{Y}_g + (\mathbf{x}_i - \bar{\mathbf{x}}_g)' \boldsymbol{\beta}_0$ with $\boldsymbol{\beta}_0 = E\left(\hat{\boldsymbol{\beta}}\right)$. Because the w_i are bounded by fixed multiples of n^{-1},

$$E\left\{\sum_{i \in A_g} w_i^2\right\} = O(G^{-1} n^{-1}). \tag{B.6}$$

Therefore, by the assumption that the variance of an estimator of a population mean is $O(n^{-1})$,

$$V\left\{\sum_{i \in A_g} w_i \left(1, e_{ig}, \mathbf{x}_i'\right)\right\} = O(G^{-1} n^{-1}), \tag{B.7}$$

$$V\left\{\sum_{i \in A_g} w_i \pi_g^{-1} R_i \left(1, e_{ig}, \mathbf{x}_i'\right)\right\} = O(G^{-1} n^{-1}). \tag{B.8}$$

Let

$$(\sum_{i\in A_g} w_i, \sum_{i\in A_g} \pi_g^{-1} w_i R_i, \sum_{i\in A_g} w_i \hat{e}_{ig}, \sum_{i\in A_g} \pi_g^{-1} w_i R_i \hat{e}_{ig}) = (\bar{z}_{g1}, \bar{z}_{g2}, \bar{e}_{g1}, \bar{e}_{g2})$$

and

$$\left(\sum_{i\in A_g} w_i \mathbf{x}_i, \sum_{i\in A_g} \pi_g^{-1} w_i R_i \mathbf{x}_i\right) = (\bar{\mathbf{x}}_{g1}, \bar{\mathbf{x}}_{g2})$$

$$\left(\sum_{i\in A_g} w_i e_{ig}, \sum_{i\in A_g} \pi_g^{-1} w_i R_i e_{ig}\right) = (\bar{u}_{g1}, \bar{u}_{g2}).$$

Since $\hat{e}_i = e_{ig} - \mathbf{x}_i' \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right)$, we have $\bar{e}_{g1} = \bar{u}_{g1} - \bar{\mathbf{x}}_{g1}' \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right)$ and $\bar{e}_{g2} = \bar{u}_{g2} - \bar{\mathbf{x}}_{g2}' \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right)$. Using (13) and (B.8), we have

$$\begin{aligned} Var(\bar{e}_{g2}) &= Var(\bar{u}_{g2})[1 + o(1)] \\ &= O\left(G^{-1}n^{-1}\right), \end{aligned} \quad \text{(B.9)}$$

and, similarly, by (B.7),

$$Var(\bar{e}_{g1}) = O\left(G^{-1}n^{-1}\right). \quad \text{(B.10)}$$

Then by (B.7), (B.8), (B.9), and (B.10),

$$\begin{aligned} (\bar{z}_{g1}, \bar{z}_{g2}, \bar{e}_{g1}, \bar{e}_{g2}) &= (\bar{z}_{Ng}, \bar{z}_{Ng}, \bar{e}_{Ng}, \bar{e}_{Ng}) \\ &+ O_p(G^{-\frac{1}{2}} n^{-\frac{1}{2}}) \end{aligned}$$

and

$$\bar{z}_{Ng}^{-1} = N_g^{-1} N = O_p(G).$$

Then

$$\hat{\theta}_{FE} - \hat{\theta}_n = \sum_{g=1}^{G} \bar{z}_{g1} \bar{z}_{g2}^{-1} (\bar{e}_{g2} - \bar{e}_{g1}) \quad \text{(B.11)}$$

and by a Taylor expansion

$$\begin{aligned} \bar{z}_{g1} \bar{z}_{g2}^{-1} (\bar{e}_{g2} - \bar{e}_{g1}) &= \left[1 + \bar{z}_{Ng}^{-1} (\bar{z}_g - \bar{z}_{g2})\right] (\bar{e}_{g2} - \bar{e}_{g1}) \\ &+ O_p(G^{1/2} n^{-1.5}) \\ &= \bar{e}_{g2} - \bar{e}_{g1} + O_p(n^{-1}). \end{aligned} \quad \text{(B.12)}$$

Because the estimator is defined to have moments,

$$\hat{\theta}_{FE} - \hat{\theta}_n = \sum_{g=1}^{G} \sum_{i\in A_g} w_i \left(\pi_g^{-1} R_i - 1\right) \hat{e}_i + O_p(Gn^{-1})$$

and

$$\tilde{\theta}_{FE} = \hat{\theta}_n + \sum_{g=1}^{G} \sum_{i\in A_g} \left(\pi_g^{-1} R_i - 1\right) w_i e_{ig}.$$

By assumption (B.2), $Gn^{-1} = o(n^{-1/2})$ and the result is established.

References

ELTINGE, J. L. and YANSANEH, I. S. (1997). Diagnostics for formation of nonresponse adjustment cells, with an application to income nonresponse in the U.S. consumer expenditure survey. *Survey Methodology*, 23, 33-40.

FAY, R. E. (1996). Alternative paradigms for the analysis of imputed survey data. *Journal of the American Statistical Association*, 91, 490-498.

FULLER, W. A. and KIM, J. K. (2002). Hot Deck Imputation for the Response Model. *Unpublished technical report*, Iowa State University.

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

KALTON, G. and KISH, L. (1984). Some efficient random imputation methods. *Communications in Statistics Part A - Theory and Methods*, 13, 1919–1939.

KIM, J. K. and FULLER, W. A. (2003) Fractional hot deck imputation. *Submitted*.

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

RAO, J.N.K., and SHAO, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika,* 79, 811-822.

SHAO, J. and STEEL, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, 94, 254-265.