VARIANCE ESTIMATION FOR NEAREST NEIGHBOR IMPUTATION WITH APPLICATION TO CENSUS LONG FORM DATA

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

Key Words: Fractional Imputation, Jackknife, Variance Estimation.

1. Introduction

Item nonresponse occurs when a sampled unit cooperates in the survey but fails to respond to some of the items. To compensate for item nonresponse, imputation can be used to estimate values for the missing items. Hot deck imputation is the imputation procedure in which the value assigned for a missing item is taken from respondents in the current sample.

Many of the hot deck imputation procedures use auxiliary variables known for both the respondents and nonrespondents to divide the sample into cells, called imputation cells. The hot deck imputation method assigns the value from a record with a response to the record with a missing value for the same cell. The record providing the value is called the donor and the record with the missing value is called the recipient. A property of hot deck imputation is that any imputed value is a known possible value of the study variable. For example, imputed values for categorical variables will also be categorical with the same number of categories as observed for the respondents.

Nearest Neighbor imputation (NNI) is a type of hot deck that is used for many surveys. Sande (1983) reviewed the general features of NNI approach and Rancourt, Sarndal, and Lee (1994) proposed a variance estimation method with NNI under a linear regression model. Recently, Fay (1999) addressed variance estimation problems in a simple situation, Chen and Shao (2000) proposed a model-based variance estimator, and Chen and Shao (2001) proposed a jackknife variance estimator that is less dependent on the assumed model compared to the other methods. The most challenging part of variance estimation problem for NNI is the absence of a general explicit imputation model. If we had an explicit model, then the NNI method might be replaced by the direct model-based method. For this reason, the methods of Rancourt, Sarndal, and Lee (1994) and Chen and Shao (2000), which are derived under explicit models, may not be suitable as a general methodology of variance estimation for NNI. Fay (1999) gave a reasonable set of assumptions but his variance estimator can lead to negative estimates for some domains. Only the method of Chen and Shao (2001) leads to a valid variance estimator under a fairly general imputation model with weak assumptions. However, their variance estimator does not apply unless the imputation classes are constructed using strata. The method of Chen and Shao (2001) is also based on the assumption that the number of imputation cells is fixed, the number of respondents is large within each imputation cell, and that the sampling fraction is negligible. Thus, their method is not applicable to the Long form Census data where the sampling fraction is sizeable.

Census 2000 Long form data are obtained from a stratified random sample and the Census long form data estimation operation uses a NNI method for handling item nonresponse. For variance estimation, Fairchild (2001) discussed options for incorporating imputation variance into long form direct variance estimates and concluded that the current methods of variance estimation after NNI cannot be applied to Long form data. The purpose of this report is to provide a variance estimation method suitable for the long form variance estimation. The method we propose in this report will produce unbiased variance estimates under the assumptions used by Fay (1999). The proposed method naturally incorporates sampling fractions and can be implemented using a replication method such as the jackknife. Before we propose a variance estimator for the NNI method, we first consider a more general variance estimation methodology for hot deck imputation. Kim and Fuller (2001, hereafter KF) established variance estimation methodology for any hot deck imputation method. Since NNI is a special case of hot deck imputation, we can modify the idea of KF for the NNI method. The cell mean model, which assumes an iid model within each cell, is replaced by the assumption of Fay (1999), wherein it is assumed that the iid model holds within a neighborhood, defined

by the nearest neighbors used in imputation. The techniques of variance estimation in KF are based on fractional imputation, that involves replicating a nonrespondents' records a number of times and imputing separately to each replicate, usually with different donors. Since the long form survey imputed only one value for each missing item, it is necessary to modify the KF procedure to make it applicable to a single imputation method.

In Section 2, we introduce the notation and assumptions used for general hot deck imputation. In Section 3, the variance estimation method of KF is reviewed and some examples are used to illustrate the ideas. In Section 4, we develop the method of KF for NNI. In Section 5 modifications for the single imputation of the U. S. Census long form are presented. Concluding remarks are in Section 6.

2. Preliminaries

2.1 Hot deck imputation

A hot deck imputation method can be described by two factors, the first of which is the way in which donors are selected for each missing item. This is determined by the distribution of $\mathbf{d} = (d_{ij};\ i \in A_R, j \in A_M)$, where A_R denotes the set of indices of the sample respondents, A_M denotes the set of indices of the sample nonrespondents , and d_{ij} is the number of times that Y_i is used as donor for Y_j. The distribution of $\mathbf{d}$ is called the *imputation mechanism*.

The second factor is the way the weight of the donor is defined for each missing item. Let w_{ij}^* be the fraction of the original weight assigned to donor i as a donor for element j. For missing item j,

$$Y_{Ij} = \sum_{i \in A_R} d_{ij} w_{ij}^* y_i \tag{1}$$

is the weighted mean of the imputed values. In the case of a single imputed values the full fraction $\left(w_{ij}^* = 1\right)$ is assigned to the donor and all other respondents are given zero weights for missing unit j. The d_{ij} are nonnegative and the sum of the imputation fractions, w_{ij}^* of the donors for a missing item should be equal to one. Thus,

$$\sum_{i \in A_R} d_{ij} w_{ij}^* = 1, \quad \forall j \in A. \tag{2}$$

A linear estimator using hot deck imputation can be written in the form

$$\hat{\theta}_I = \sum_{i \in A_R} \left(\sum_{j \in A} d_{ij} w_j w_{ij}^* \right) y_i \equiv \sum_{i \in A_R} \alpha_i y_i. \tag{3}$$

The sum of $w_j w_{ij}^*$ over all recipients for which i is a donor (including the donor for itself), denoted by α_i, is the total weight of donor i. Note that $\alpha_i \geq w_i$ for $i \in A_R$, because the final weight of a unit is increased if the unit is used as a donor. If responding unit i is not used as a donor, except for itself, then $\alpha_i = w_i$.

2.2 Cell mean model

Assume that the finite population U is made up of G imputation cells. Let n_g be the number of sample elements in imputation cell g and let r_g, $r_g > 0$, be the number of respondents in imputation cell g. Within cell g, $g = 1, 2, \cdots, G$, the elements in the finite population are a realization of independently and identically distributed random variables with mean μ_g and variance σ_g^2. Thus,

$$Y_i \overset{i.i.}{\sim} \left(\mu_g, \sigma_g^2\right), \quad \forall i \in U_g \tag{4}$$

where U_g denotes the set of indices for the g-th imputation cell in the population U and $\overset{i.i.}{\sim}$ is the abbreviation for independently and identically distributed. We call the model (4) the cell mean model.

The distribution of Y in the sample is determined by the sampling mechanism and by the distribution of the vector Y. If there is no dependence of the distribution of Y on the sampling mechanism, the sampling mechanism is said to be ignorable. That is, a sampling mechanism is ignorable if the conditional distribution of Y for the sample is equal to the marginal distribution of Y. Similarly, the response mechanism is said to be ignorable if the conditional distribution of Y for the respondents is the same as the conditional distribution of Y for the sample.

We assume the sampling mechanism and the response mechanism are ignorable under the cell mean model (4). Then, the cell mean model holds for the respondents and for the nonrespondents in the sample. Thus,

$$Y_i \mid (A, A_R) \overset{i.i.}{\sim} \left(\mu_g, \sigma_g^2\right), \quad \forall i \in U_g \tag{5}$$

The expression in (4) is the marginal distribution of the population vector $\mathbf{Y}$, while the expression in (5) is the conditional distribution of the population vector $\mathbf{Y}$ given the realized sample and the realized respondents.

The imputation mechanisms allowed are quite general, including with-replacement hot deck and without-replacement hot deck procedures. The only restrictions on the imputation mechanism are:

(I.1) The distribution of $\mathbf{d}$ is independent of $\mathbf{Y}$ and depends only on $(r_1, r_2, \cdots, r_G)$ and $(n_1, n_2, \cdots, n_G)$.

(I.2) For $i \in A_R$ and $j \in A_M$,

$$Pr\left(d_{ij} > 1 \mid i \in U_h, j \in U_g\right) = 0 \quad \text{if } h \neq g$$
$$> 0 \quad \text{if } h = g.$$

The two conditions make the distribution of the respondents unchanged after imputation. Thus, the resulting distribution of the respondents after imputation is

$$Y_i \mid (A, A_R, \mathbf{d}) \overset{i.i.}{\sim} \left(\mu_g, \sigma_g^2\right), \quad \forall i \in U_g \tag{6}$$

2.3 Variance of hot deck imputation estimator

Under the assumptions discussed in Section 2.2, KF show that the variance of the hot deck imputation estimator (3) is

$$Var\left(\hat{\theta}_I\right) \tag{7}$$
$$= Var\left(\sum_{g=1}^{G}\sum_{i\in A_g} w_i\mu_g\right) + E\left(\sum_{g=1}^{G}\sum_{i\in A_{Rg}} \alpha_i^2\sigma_g^2\right),$$

where $A_g = A \cap U_g$ is the set of indices for the g-th imputation cell in the sample and $\alpha_i = \sum_{j\in A} w_j w_{ij}^* d_{ij}$ is the total weight of donor i after hot deck imputation. The distribution used to define the variance in (7) is the joint distribution of the cell mean model, sampling mechanism, response mechanism, and imputation mechanism.

3. Variance estimation after hot deck imputation

To consider variance estimation, let a replication variance estimator for the complete sample be

$$\hat{V}\left(\hat{\theta}\right) = \sum_{k=1}^{L} c_k\left(\hat{\theta}^{(k)} - \hat{\theta}\right)^2, \tag{8}$$

where $\hat{\theta}^{(k)}$ is the k-th estimate of θ_N based on the 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 $\hat{\theta}$ is a linear estimator of the form (1), the k-th replicate of can be written $\hat{\theta}^{(k)} = \sum_{i\in A} w_i^{(k)} y_i$, where $w_i^{(k)}$ denotes the replicate weight for the i-th unit of the k-th replicate.

If we treat the imputed values as if they are true values and apply the standard replication variance estimator in (9), then the naive variance estimator can be expressed as

$$\hat{V}_1\left(\hat{\theta}_I\right) = \sum_{k=1}^{L} c_k\left(\hat{\theta}_{I1}^{*(k)} - \hat{\theta}_I\right)^2,$$

where $\hat{\theta}_{I1}^{(k)} = \sum_{i\in A_R} \alpha_{i1}^{(k)} y_i$ with $\alpha_{i1}^{(k)} = \sum_{j\in A_R} w_j^{(k)} w_{ij}^* d_{ij}$. Kim and Fuller (1999) showed that, if the complete sample variance estimator in (8) is design unbiased for the variance of $\hat{\theta}_n$, then the naive variance estimator applied to the imputed data set satisfies

$$E\left\{\hat{V}_1\left(\hat{\theta}_I\right)\right\} = Var\left(\sum_{g=1}^{G}\sum_{i\in A_g} w_i\mu_g\right) \tag{9}$$
$$+ \sum_{k=1}^{L} E\left\{\sum_{g=1}^{G}\sum_{i\in A_{Rg}} c_k\left(\alpha_{i1}^{(k)} - \alpha_i\right)^2\sigma_g^2\right\}.$$

Thus, comparing (9) with (7), we can see that the naive variance estimator unbiasedly estimates the first component of the total variance in (7) but does not necessarily unbiasedly estimate the second component of the total variance. The bias of the naive variance estimator is

$$Bias\left(\hat{V}_I\right) = E\left\{\sum_{g=1}^{G}\sum_{i\in A_{Rg}}\left[\sum_{k=1}^{L}\left(\alpha_{i1}^{(k)} - \alpha_i\right)^2 - \alpha_i^2\right]\sigma_g^2\right\}.$$

The bias of the naive variance estimator comes from the fact that the contribution of donor i to the total variance (α_i^2) is different from the contribution of donor i to the expectation of the naive variance estimator ($\sum_{k=1}^{L} c_k\left(\alpha_{i1}^{(k)} - \alpha_i\right)^2$). In that sense, the replicate weight $\alpha_{i1}^{(k)} = \sum_{j\in A} w_j^{(k)} w_{ij}^* d_{ij}$ is not the proper weight to use in estimating the imputation variance. To correct this, we compute replicates of the imputation fractions that account for the realized total weight due to imputation. Thus, if we can construct a replication method defined by

$$\hat{\theta}_I^{(k)} = \sum_{i\in A_R} \alpha_i^{(k)} y_i = \sum_{i\in A_R}\left(\sum_{j\in A} w_j^{(k)} w_{ij}^{*(k)} d_{ij}\right) y_i$$

so that

$$\sum_{i\in A_{Rg}}\sum_{k=1}^{L}\left(\alpha_{i1}^{(k)} - \alpha_i\right)^2 = \sum_{i\in A_{Rg}} \alpha_i^2, \quad g = 1, 2, \cdots, G, \tag{10}$$

then the second component of the total variance in (7) will be unbiasedly estimated. In addition to (10), we also require

$$\sum_{i\in A_R} d_{ij} w_{ij}^{*(k)} = 1, \quad \forall j \in A. \tag{11}$$

Note that the replicates of the imputation fractions mimic the behavior (2) of original imputation fractions.

A sufficient condition for (10) is

$$\sum_{i\in A_{Rg}} c_k \left(\alpha_i^{(k)} - \alpha_i\right)^2 - \sum_{i\in A_{Rg}} c_k \left(\alpha_{1i}^{(k)} - \alpha_i\right)^2 = \alpha_k^2 - \sum_{i=1}^{L} c_i \left(\alpha_{1k}^{(k)} - \alpha_k\right)^2 \quad (12)$$

where $\alpha_{1i}^{(k)} = \sum_{j\in A} w_j^{(k)} w_{ij}^* d_{ij}$ is the replicate total weight for the naive variance estimator. The left side of equality (12) is the difference between the contribution of k-th replicate to the expected value of the proposed variance estimator and the contribution of k-th replicate to the expected value of the naive variance estimator. The right side of equality (12) is the difference between the contribution of unit k to the conditional variance and the contribution of unit k to the expected value of the naive variance estimator.

To construct an unbiased estimator of the variance we derive replicates that satisfy both (11) and (12). A natural starting place is to consider replicates constructed by removing all of the imputed values associated with a deleted respondent and increasing the weights of the other donors. By slightly modifying this procedure with an adjustment for each replicate we can construct an unbiased estimator of the variance. For a fractionally imputed procedure with M distinct donors, replicates satisfying (11) are

$$w_{ij}^{*(k)} = \begin{cases} w_{ij}^* - \delta_k & \text{if } d_{kj} = 1 \text{ and } k = i \\ w_{ij}^* + (M-1)^{-1}\delta_k & \text{if } d_{kj} = 1 \text{ and } k \neq i \\ w_{ij}^* & \text{otherwise,} \end{cases} \quad (13)$$

where δ_k are calculated to satisfy (12). Replicates that satisfy (12) satisfy

$$c_k \left\{\alpha_{1k}^{(k)} - \alpha_k - \delta_k \sum_{j\in A_M} w_j^{(k)} d_{kj}\right\}^2 + \sum_{\substack{i\in A_{Rg} \\ i\neq k}} c_k \left\{\alpha_{1i}^{(k)} - \alpha_k + \frac{\delta_k}{M-1} \sum_{j\in A_M} w_j^{(k)} d_{kj} d_{ij}\right\}^2 - \sum_{i\in A_{Rg}} c_k \left\{\alpha_{1i}^{(k)} - \alpha_i\right\}^2 = \alpha_k^2 - \sum_{i=1}^{L} c_i \left\{\alpha_{1k}^{(i)} - \alpha_k\right\}^2.$$

Thus, for mutually exclusive imputation cells, a δ_k can be determined by solving a quadratic equation of δ_k. If such δ_k exists, the variance estimator will be unbiased. In contrast, the method of Chen and Shao (2000, 2001) is based on asymptotic theory, where the number of respondents in every cell must be large enough to apply large sample theory.

Under simple random sampling, for sufficiently large sample size n, determining equation (12) can be simplified as

$$(1+\delta_k d_k)^2 + \frac{d_k^2 \delta_k^2}{M-1} = \left(1 + \frac{d_k}{M}\right)^2. \quad (14)$$

By the mean value theorem, we can show that the solution to (14) satisfies $0 < \delta_k < M^{-1}$, which guarantees nonnegative replicates of w_{ij}^*.

4. Application to NNI method

The variance estimation method described in Section 3 can be applied for any hot deck imputation method. Fay (1999) assumed

$$\begin{aligned} E_\zeta(y_b) &= E_\zeta\left(y_{nnt(b)}\right) \\ Var_\zeta(y_b) &= Var_\zeta\left(y_{nnt(b)}\right) \end{aligned} \quad (15)$$

and assumed the y-variables in the neighborhood to be uncorrelated, where $nnt(b)$ is the index for neighbor t of unit b. We make the stronger assumption that

$$y_b \overset{i.i.}{\sim} \left(\mu_i, \sigma_i^2\right) \quad \forall b \in B_i \quad (16)$$

where B_i is the set of neighborhoods of element i based on the neighbor designation. We call B_i the nearest neighbor cell or the neighborhood of y_i. The cells B_i can overlap, and the validity of the KF method does not require the cells to be disjoint. Only (10) and (11) are required. Then (16) is essentially a cell mean model, where B_i is a cell. Thus, we can apply the method described in Section 3 and form replicates to satisfy (10).

If Y_i appears in more than one neighborhood we assume the variance is the same in all neighborhoods containing Y_i. Then equation (7) becomes

$$Var\left(\hat{\theta}_n\right) = Var\left\{\sum_{i\in A} w_i \mu_i\right\} + E\left\{\sum_{i\in A_R} \alpha_i^2 \sigma_i^2\right\}. \quad (17)$$

A replicate variance estimator will be unbiased if

$$\sum_{k=1}^{L} c_k \sum_{t\in B_i} \left(\alpha_t^{(k)} - \alpha_t\right)^2 = \sum_{t\in B_i} \alpha_t^2, \quad \forall i, \quad (18)$$

where

$$\alpha_t^{(k)} = \sum_{j\in D_t} w_j^{(k)} w_{tj}^{*(k)}$$

and D_i is the set of recipients for which i is a donor. Equation (18) is analogous to equation (10).

To discuss construction of the replicate fraction $w_{tj}^{*(k)}$, let $\mathcal{P}_k$ be the set of indices of elements deleted,

or that have their weights reduced, in the k-th replicate. For $t \in B_i$, the replication fraction $w_{tj}^{*(k)}$ is

$$w_{tj}^{*(k)} = \begin{cases} w_{tj}^* - \delta_k & \text{if } M_{jk} > 0 \text{ and } t \in \mathcal{P}_k \\ w_{tj}^* + C_{jk}\delta_k & \text{if } t \notin \mathcal{P}_k \\ w_{tj}^* & \text{otherwise,} \end{cases} \tag{19}$$

where $M_{jk} = \sum_{t \in \mathcal{P}_k \cap A_R} d_{ij}$ is the number of donors in $\mathcal{P}_k$ for missing unit j and $C_{jk} = (M - M_{jk})^{-1} M_{jk}$. Note that (19) is essentially the same as (13). The only difference is that B_i is treated as a cell. Assuming the $\mathcal{P}_k$ to be mutually exclusive and exhaustive, the δ_k can be calculated by

$$\begin{aligned} &\sum_{t \in B_i} c_k \left(\alpha_t^{(k)} - \alpha_t\right)^2 - \sum_{t \in B_i} c_k \left(\alpha_{1t}^{(k)} - \alpha_t\right)^2 \\ = &\sum_{t \in \mathcal{P} \cap B_i} \left\{ \alpha_t^2 - \sum_{s=1}^{L} c_s \left(\alpha_{1t}^{(s)} - \alpha_t\right)^2 \right\} \end{aligned} \tag{20}$$

where $\alpha_{1t}^{(k)} = \sum_{j \in D_t} w_j^{(k)} w_{tj}^*$. By summing both sides of (20) over replicates it can be shown that (20) implies (18).

5. Application to the U. S. Census Long Form

To apply the method of KF to the Census Long Form data we extend the definition of fractional imputation to include single imputation. We assume that the neighborhood has M donors, but that $w_{ij}^* = 1$ if $d_{ij} = 1$ and $w_{ij}^* = 0$ if $d_{ij} = 1$. That is, single imputation becomes a special case of fractional imputation. Then, we can apply the KF method to a single imputation and the suggested replicate fraction for a missing unit is (13). Thus, if M = 2 deleting a first nearest neighbor will reduce the fractional weight of the donor and increase the fractional weight of the second nearest neighbor. But, deleting a second nearest neighbor will have no effect on the replicate fractional weight of donors. The proposed method for variance estimation uses fractional imputation even though a single imputation was used for point estimation. Fractional imputation was originally proposed to reduce the variance due to imputation and this is not achieved with a single imputation.

6. Concluding Remarks

In this paper, we examined some properties of the hot deck imputation estimator and a variance estimator based on the KF paper. We showed that modifications of the KF method make it applicable to NNI data. The proposed replication variance estimation method is exactly unbiased under fairly reasonable assumptions, and covers nonignorable sampling fractions. The proposed variance estimator is also applicable to estimating means for several items with different missing patterns. A disadvantage of the proposed method is that the computation can be cumbersome for a large data set, because it requires solving quadratic equations. For the Long Form Census data, we recommend using the proposed method to evaluate the approximate prediction variance estimation method already in place. We have demonstrated that the proposed method can be implemented for the jackknife method. Implementation for Fay's successive difference replication method (Fay and Train, 1995) is also possible. This memo, does not include an empirical investigation of the proposed method. In the KF paper, extensive simulations show the superiority of the proposed method over the multiple imputation method. We plan to include empirical findings in the next version of this memo.

Reference

Chen, J. and Shao, J. (2000). Nearest neighbor imputation for survey data. Journal of Official Statistics. 16, 113-132.

Chen, J. and Shao, J. (2001). Jackknife variance estimation for nearest-neighbor imputation. Journal of American Statistical Association. 96, 260-269.

Fairchild, L. (2001). Options for incorporating imputation variance into long form direct variance estimation for census 2000. DSSD Census 2000 memorandum series. May 31.

Fay, R.E. (1999). Theory and application of nearest neighbor imputation in Census 2000. Proceedings of the Section on Survey Research Methods (pp. 112-121). Alexandria, VA: American Statistical Association.

Fay, R.E. and Train G.F.(1995). Aspects of survey and model-based postcensal estimation of income and poverty characteristics for states and counties. Proceedings of the Section on Government Statistics (pp. 154-159). Alexandria, VA: American Statistical Association.

Kim, J.K., and Fuller, W.A. (1999). Jackknife variance estimation after hot deck imputation. Proceedings of the Section on Survey Research Methods (pp. 825-830). Alexandria, VA: American Statistical Association.

Kim, J. and Fuller, W (2001). Inference procedures for hot deck imputation. Submitted.

Rancourt, E., Sarndal, C.E, and Lee, H. (1994). Estimation of the variance in the presence of nearest neighbor imputation. Proceedings of the Section on Survey Research Methods (pp. 888-893). Alexandria, VA: American Statistical Association.

Sande, I. (1983). Hot-deck imputation procedures, in Incomplete Data in Sample Surveys, Vol 2. (pp. 339-349). Academic Press, New York.

Sarndal, C.E. (1992). Methods for estimating the precision of survey estimates when imputation has been used. Survey Methodology, 18, 241-252.