Imputation using response probabilities

Jae-Kwang Kim[1], Hyeon-Ah Park[2], Jongwoo Jeon[2]

ABSTRACT

In this paper, we propose a class of imputed estimators using response probability. The proposed estimator can be justified under the response probability model and thus is robust against the failure of the assumed imputation model. We also propose a variance estimator that is justified under the response probability model.

KEY WORDS. *Nonresponse, Survey sampling, Variance estimation.*

1. Introduction

Imputation is a commonly used method of compensating for item nonresponse in sample surveys. 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. Imputation model refers to the assumption about the variables collected in the survey and the relationship among these variables. Often the imputation model is quite difficult to verify from a single data set because of the missing value of the study variable. Another approach, called response probability model approach, is also commonly adopted in the analysis of missing data. Response probability model refers to the assumptions about the probability of obtaining a response from the sampled unit for the item. One of the commonly used response probability model is the cell response model, where the responses are assumed to be uniform within the imputation cell. Rao and Shao (1992) and Shao and Steel (1999) discuss inferences of the imputed estimator under the cell response model. However, for the other response models such as logistic response model, imputation methods incorporating the response probability model are relatively underdeveloped, although analyses incorporating the response probability model are quite common in the non-imputation context. Examples include Rosenbaum (1987), Robins et al (1994), and Lipsitz et al (1999).

[1]Department of Statistics, Hankuk University of Foreign Studies, Yongin, Kyungki-Do 449-791.
[2]Department of Statistics, Seoul National University, Seoul, 151-742.

The purpose of this paper is to provide an imputation methodology of combining the imputation model and the response model. In section 2, an imputation method is proposed that can be justified under the two approaches. Thus, the resulting estimator is "doubly" protected against the failure of the assumed model. In Section 3, the proposed method is applied to the ratio imputation model. In Section 4, a variance estimator that is justified under the response model is proposed. Concluding remarks are made in Section 5.

2. Basic setup

Let $\hat{\theta}_n$ be an estimator of the population parameter θ_N based on the sample and of the form $\hat{\theta}_n = \sum_{i=1}^{n} w_i Y_i$, where w_i is the sampling weight and Y_i is the study variable of the i-th element in the sample of size n. We assume that

$$E_D\left(\hat{\theta}_n\right) = \theta_N. \tag{1}$$

where the expectation is taken with respect to the sampling mechanism. Under nonresponse, we define the response indicator function of Y_i

$$R_i = \begin{cases} 1 & Y_i \text{ responds} \\ 0 & Y_i \text{ does not respond} \end{cases}, \quad i = 1, 2, \cdots, n,$$

and its expectation $\pi_i = Pr\left(R_i = 1 \mid i\right)$.

If we define Y_i^* to be the imputed value of Y_i, then the estimator of θ_N based on the imputed values can be written

$$\hat{\theta}_I = \sum_{i=1}^{n} w_i \left\{R_i Y_i + \left(1 - R_i\right) Y_i^*\right\}. \tag{2}$$

The imputed values usually satisfies

$$E_\zeta\left(Y_i^*\right) = E_\zeta\left(Y_i\right), \quad i = 1, 2, \cdots, n, \tag{3}$$

where the expectation in (3) is with respect to the conditional distribution of Y given the respondent status. The model involved in (3) is called *imputation model*. Assumption (3) implies $E_\zeta\left(\hat{\theta}_I - \hat{\theta}_n\right) = 0$, and so, by (1),

$$E_D E_\zeta\left(\hat{\theta}_I - \theta_N\right) = 0. \tag{4}$$

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.

If we know the response probability π_i, then we can use the response probability to relax the assumption (3). The proposed estimator is

$$\hat{\theta}_{Id} = \sum_{i=1}^{n} w_i Y_i^* + \sum_{i=1}^{n} w_i \pi_i^{-1} R_i \left(Y_i - Y_i^*\right). \tag{5}$$

Note that

$$\hat{\theta}_{Id} - \hat{\theta}_n = \sum_{i=1}^{n} w_i \left(\pi_i^{-1} R_i - 1\right) \left(Y_i - Y_i^*\right). \tag{6}$$

If $E(\pi_i^{-1} R_i) = 1$, under the assumptions discussed in Section 3, the right side of (6) is asymptotically negligible. Thus, the proposed estimator in (5) is approximately unbiased for θ_N under the response mechanism, regardless of whether the imputation model holds or not.

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

$$\sum_{i=1}^{n} w_i \left(\pi_i^{-1} - 1\right) R_i \left(Y_i - Y_i^*\right) = 0. \tag{7}$$

Hence, condition (7) suggests a way of constructing an imputed estimator. In the next section, we illustrate how to construct an imputed estimator satisfying (7) under the ratio imputation model.

3. Application to ratio imputation model

Suppose that we have a completely observed auxiliary variable x_i for the i-th unit in the sample. A commonly used imputation model is the ratio imputation model

$$E_\zeta \left(Y_i\right) = x_i \gamma. \tag{8}$$

Under the ratio imputation model, 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), the choice of $\hat{\gamma}$ was

$$\hat{\gamma} = \left\{ \sum_{i=1}^{n} w_i R_i x_i \right\}^{-1} \sum_{i=1}^{n} w_i R_i Y_i. \tag{9}$$

In our new approach, a choice of $\hat{\gamma}^*$ satisfying (7) is

$$\hat{\gamma}^* = \left\{ \sum_{i=1}^{n} w_i \left(\pi_i^{-1} - 1\right) R_i x_i \right\}^{-1} \sum_{i=1}^{n} w_i \left(\pi_i^{-1} - 1\right) R_i Y_i, \tag{10}$$

which reduces to (9) under the uniform response mechanism, where the π_i are a constant. Thus, the imputed estimator $\hat{\theta}_I$ using $\hat{\gamma}^*$ in (10) is algebraically equivalent to $\hat{\theta}_{Id}$ in (5).

The following theorem shows that the the proposed estimator is asymptotically unbiased without assuming the imputation model. The reference distribution in (15) and (16) is the joint distribution of the sampling mechanism and the response mechanism.

Theorem 3.1 *Let $\hat{\theta}_n$ be a design unbiased complete sample estimator for the population parameter θ_N. Assume a sequence of finite populations with finite 4-th moments of (x_i, Y_i) as defined in Isaki and Fuller (1982). Assume the sampling mechanism satisfy*

$$K_1 < \max_i n w_i < K_2 \tag{11}$$

and

$$nVar\left(\hat{\theta}_n\right) > K_3 \tag{12}$$

for some nonnegative constants K_1 ,K_2, and K_3, uniformly in n. Assume that response mechanism satisfy

$$K_4 < \pi_i, \tag{13}$$

for some nonnegative constants K_4, and

$$Pr\left(R_i = 1, R_j = 1\right) = Pr\left(R_i = 1\right) Pr\left(R_j = 1\right), \quad \forall i \neq j. \tag{14}$$

Then, the imputed estimator of the form (2) with $Y_i^* = x_i\hat{\gamma}^*$ satisfy

$$E\left(\hat{\theta}_{Id}\right) = \theta_N + o\left(n^{-1/2}\right) \tag{15}$$

and

$$V\left(\hat{\theta}_{Id}\right) = V_D\left(\hat{\theta}_n\right) + E_D\left[\sum_{i=1}^{n} w_i^2\left(\pi_i^{-1} - 1\right)\left(Y_i - x_i\gamma^0\right)^2\right] + o(n^{-1}), \tag{16}$$

where $\gamma^0 = E_{DR}\left(\hat{\gamma}^*\right)$ and the subscript D denote the distribution over the sampling mechanism.

4. Variance Estimation

We now consider the variance estimation of the imputed estimator satisfying (7) under the response model approach. We adopt replication method such as jackknife for variance estimation. Replication variance estimator is popular because it can be easily extended to the variance estimation for non-linear statistics.

Under complete response, let a replication variance estimator be

$$\hat{V} = \sum_{k=1}^{L} c_k\left(\hat{\theta}_n^{(k)} - \hat{\theta}_n\right)^2, \tag{17}$$

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 , the k-th replicate of $\hat{\theta}_n$ can be written $\hat{\theta}_n^{(k)} = \sum_{i=1}^n w_i^{(k)} Y_i$, where $w_i^{(k)}$ denotes the replicate weight for the i-th unit of the k-th replication.

Under nonresponse, we propose a variance estimator for the imputed estimator of the form in (5) using the replication method in (17). The proposed replication variance estimator is

$$\hat{V}_d = \sum_{k=1}^{L} c_k \left(\hat{\theta}_{Id}^{(k)} - \hat{\theta}_{Id} \right)^2, \tag{18}$$

where

$$\hat{\theta}_{Id}^{(k)} = \sum_{i=1}^{n} w_i^{(k)} Y_i^{*(k)} + \sum_{i=1}^{n} w_i^{(k)} \pi_i^{-1} R_i \left(Y_i - Y_i^{*(k)} \right) \tag{19}$$

and $\hat{\theta}_{Id}$ is defined in (5). The $Y_i^{*(k)}$ is a replicated version of Y^* satisfying

$$\sum_{i=1}^{n} w_i^{(k)} \left(\pi_i^{-1} - 1 \right) R_i \left(Y_i - Y_i^{*(k)} \right) = 0. \tag{20}$$

Note that condition (20) for the replicates is similar to condition (7) for the original estimator.

In the following theorem, we show the consistency of the proposed jackknife variance estimator under the response probability model.

Theorem 4.1 *Let the assumptions of Theorem 3.1 hold. Let the replication variance estimator for the complete sample be of the form (17). Assume that*

$$\max_k c_k^{-1} = O(L) \tag{21}$$

and

$$E_D \left\{ \left[c_k \left(\hat{\theta}^{(k)} - \hat{\theta} \right)^2 \right]^2 \right\} \quad < \quad C_\theta L^{-2} \left[V_D \left(\hat{\theta} \right) \right]^2 \tag{22}$$

for all k and for some constant C_θ. Assume

$$E_D \left\{ \left[\hat{V}/V(\hat{\theta}) - 1 \right]^2 \right\} \quad = \quad o(1) \tag{23}$$

for any y with bounded fourth moments. We also assume that the sampling fraction is negligible.

$$N^{-1} n = o(1). \tag{24}$$

Then the proposed jackknife variance estimator defined in (18) satisfies

$$n\left\{\hat{V}_d - V\left(\hat{\theta}_{Id}\right)\right\} = o_p(1). \tag{25}$$

5. Concluding remarks

A class of imputed estimator using the response probability is proposed. The proposed estimator are asymptotically unbiased even under the failure of the assumed imputation model, as long as the assumed response probabilities are true. Variance estimation using the replication method is also proposed and its asymptotic properties are presented. The proposed estimator also shows good finite sample properties in the simulation. Asymptotic properties of the imputed estimator using the estimated response probability are not discussed here and will be presented somewhere else.

References

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

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

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

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

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

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

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