Proceedings of the Autumn Conference in 2001
and the 30th Anniversary of the Korean Statistical Society

# Multiple imputation inference for stratified random sample with nonignorable nonresponse

**Minwoong Shin[1], Sangeun Lee[2], Sungchul Lee[3], Juyoung Lee[4]**


## Abstract

In general, the imputation problems which are caused from survey nonresponse have been studied for being based on ignorable cases. However the model based approach can be applied to survey with nonresponse suspected of being nonignorable. Here in this study, we will make the nonresponse for nonignorable into ignorable cell using adjustment cell approach, then we can applied the ignorable nonresponse method. For data sets of each nonresponse cells are simulated from normal distribution.


keywords :Multiple imputation, Nonignorable nonreponse, Adjustment cells.

## 1. Introduction

Multiple imputation refers to the procedure of replacing each missing value by a vector of $M \geq 2$ imputed values. When the M sets of imputations are repeated random draws under one model for nonresponse, the M complete data inferences can be combined to form one inference.

The assumption of ignorable nonresponse is that respondents and nonrespondents with the same values of recorded variables do not differ sysmatically on the values of variables missing for the nonrespondents; that is, the missing values are assumed to be MAR. By using nonignorable models for nonresponse, the model-based approach can be applied to surveys with nonresponse suspected of being nonignorable.

An important issue that arises with nonignorable nonresponse is whether the objective of the statistical analysis is first, to provide a single valid inference or secondly, to display

1) Department of Statistics, Hankuk University of Foreign Studies, Kyonggi-do, Korea, 449-791
2) Department of Applied Information Statistics Kyonggi University, Kyonggi-do, Korea, 442-760
3) Department of General Education, Namseoul University
4) Division of Epidemiology, Department of Infectious Disease, National Institute of Health, #5 Nokbun-dong, Eunpyung-gu, Seoul, Korea, 112-701

sensitivity of conclusions to nonignorable nonresponse by providing a variety of inference each which is valid under an assumed model for nonresponse.
Obviously, the objective if obtaining one valid inference is generally very demanding, since common practice there is no hard evidence from which or specify one correct model for nonresponse. Although less satisfying than providing one valid inference, displaying sensitivity through a variety if conditionally valid inference is a much more realistic goal.

## 2. A nonignorable adjustment cell model

Consider inference for a population mean $\overline{Y}$ from a stratified random semple with nonignorable missing-data. With complete data inference for $\overline{Y}$ would be made using the statement.

$$(\overline{Y} - \sum_{h=1}^{H} P_h \overline{y}_h) \sim N\left[ 0 , (\sum_{h=1}^{H} P_h^2 (n_h^{-1} - N_h^{-1}) S_h^2) \right] \tag{2.1}$$

Suppose that ajustment cells are formed so that in cell j the distribution of Y is the same for nonrespondents and for respondents and is normal with mean $\mu_{jh}$ and variance $\sigma_{jh}^2$ in the ***hth*** strate. However, the variable defining adjustment cells is recorded only for respondents in the sample.
Suppose that nonresponse to a survey is ignorable within income categories and income is recorded for respondents but not for nonrespondents. Then we know the number of respondents $(m_{jh})$ in cell j but do not know the number of nonrespondents $(n_{jh} - m_{jh})$
If the sample sizes $n_{jh}$ in the cell j were known, we could create multiple inputation for the nonrespondents.

Tabel 1. The sample cross-classified by response and adjustment cell variables known only for respondents

| | | Response indicator R | | |
|---|---|---|---|---|
| | | $R_i = 1$ | $R_i = 0$ | Total |
| Adjustment cell | 1 | $m_1$ | $(n_1 - m_1)$ | $(n_1)$ |
| | ⋮ | ⋮ | ⋮ | ⋮ |
| | j | $m_j$ | $(n_j - m_j)$ | $(n_j)$ |
| | ⋮ | ⋮ | ⋮ | ⋮ |
| | J | $m_J$ | $(n_J - m_J)$ | $(n_J)$ |
| | Total | $m$ | $n - m$ | $n$ |

We consider a method based on an implicit model, which is called the approximate Bayesian Bootstrap by Rubin and Schenker(1986).

For $l = 1, \cdots, M$carry out the following steps independently : For each stratum, first create $n_{jh}$ possible values of Y by drawing $n_{jh}$ values at random with replacement from the $m_{jh}$ observed values of Y in the cell j and second draw the $n_{jh} - m_{jh}$missing values of Y at random with replacement from these $n_{jh}$ values.

Results in Rubin(1987) can be used to show that this method is proper for large M.

## 3. Multiple imputation inference for stratified random sample

Since the multiple imputations are drawn from a predictive distribution, an intuitive method for creating such imputation is the hot deck, which draws the nonrespondents' values at random from the respondents' values in the same stratum.

Now suppos that only $m_h$ of the $n_h$ units in stratum $h$ are respondents. With multiple imputation, each of the $\sum_h (n_h - m_h)$ missing ucits would have M imputations, these by, creating M completed data sets and M values of the stratum means and variances say, $\bar{y}_{h(l)}$ and $S^2_{h(l)}$, $l = 1, \cdots, M$ The method of imputation is described in section 2.

Note that $\sum_j m_{jh} = m_h$and $\sum_j n_{jh} = n_h$

The multiple imputation estimate $\hat{\bar{Y}}$ is the average of the M complete-data estimates of $\bar{Y}$,

$$\hat{\bar{Y}} = \sum_{l=1}^{M} \frac{\sum_{h=1}^{L} P_h \bar{y}_{h(l)}}{M} \tag{3.1}$$

The variability associated with $\hat{\bar{Y}}$ is the sum of the two components displayed in (3.2),

$$\frac{\sum_{l=1}^{M} [\sum P_h^2 (n_h^{-1} - N_h^{-1}) S_{hl}^2]}{M} + \frac{M+1}{M} \frac{\sum_{l=1}^{M} (\sum_{h=1}^{H} P_h \bar{y}_{h(l)} - \hat{\bar{Y}})^2}{M-1} \tag{3.2}$$

## 4. Discussion

We simulate the data as nonignorable nonresponse data using different response rate at each strata and also to be normally distributed. And at each strata, we are forming the adjustment cell to be ignorable nonresponse. We may use the regression imputation in place of ABB in the adjustment calls.

Up to this point, we are handling the simulated data. However, in practical, it is hard to form such adjustment cells. Therefore, that may realized that the ways of forming the adjustment cell to be ignorable nonresponse will be a good further research at each its own real data.

# Reference

1. Little.R.J.A. and Rubin.D.B.(1987). Statistical Analysis with missing data. John Wiley & Sons. Vol. No. , page
2.Rubin.D.B.(1986).Statistical matching using file concatenation with adjusted weights and Multiple Imputation.J. Business Econ..Statist. 4, 87-94
3.Rubin.D.B and schenker.N(1987). Multiple imputation for nonresponse surveys. New York: Wiley.
4.Rubin.D.B. and schenker.N(1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. J. Amer. Statist. Assoc. 8,. 336-374