On partial nonresponse situations:the hot deck imputation method

Gabriella Schoier
Università di Trieste, Dipartimento di Scienze Economiche e Statistiche
Piazzale Europa, 1
Trieste, Italia
e-mail: gabriellas@econ.univ.trieste.it

1. Introduction

Dealing with data files statisticians often have to consider the problem of missing data due both to unit nonresponse (complete nonresponse) and item nonresponse (partial nonresponse).

As it regards item nonresponse (see e.g. Kalton and Kasprzyk (1986)) different forms of imputation (deterministic or stochastic) are often used (see e.g. Little and Rubin (1987), Cicchitelli, Herzel, Montanari (1992)).A very well known technique used to fill in values for these missing values is the hot deck imputation (see Ford (1983)).

In this paper we consider the problem of variance estimation from imputed survey data under the hot deck method.

2. The Hot Deck Method

Hot deck imputation is commonly used for item non response as it has some advantages: it preserves the distribution of item values, it permits the use of the same sample weight for all items and results obtained from different analyses are consistent with one another.

The main principle of the hot deck metod is using the current data (donors) to provide imputed values for records with missing values. The procedure through which we find the donor that matches the record with missing values is different according to the particular techniques used.

The matching process is carried out using the so called filter variables, records match if they have the same values on the filter variables. A possibility is to take the value for the missing response from a respondent to the current survey. Other hot deck imputation methods include distance function matching or nearest neighbor imputation in which a nonrespondent is assigned the item value of the nearest neighbor.

Consider single imputation and multiple imputation for the hot deck method for the situation with a single imputation class.

Given a sample respondents A of size n, suppose that the values of the item y are observed only for a subset A_r of size r, a simple random sample of size (n-r) is selected with replacement from A_r and the associated item values y_i , $i \in A_r$ are used as donors . In practice the imputed values are treated as if they are true values, and the variance estimates are computed using standard formulas for a specified sample design even if this may cause understimation of the true variance of the estimates.

Alternatively (see Rubin (1978), Little and Rubin (1987)) the multiple imputation approach gives m (m $\geq$ 2) imputed values for each nonrespondent, if we consider the case in which the data are missing at random we can compute m different estimates $\bar{y}_{MIl}$ (l=1,…,m) of the population mean, the multiple estimate $\bar{y}_{MI}$ is given by

$$(1) \quad \overline{y}_{MI.} = \sum_{l=1}^{m} \overline{y}_{MI_l} \Big/ m$$

Alternative approaches have been proposed by Rao and Shao (1992), Rao (1996) that considered a modification to the standard design stratified jacknife variance formula to obtain suitable estimates for a data set of missing data uncertanty on the base of a single imputation, Fay (1996) extended Rao and Shao results to fractionally weighted imputation and multiple imputation .

In this paper we deal with the variance estimation from imputed survey data under this imputation method also considering simulation studies.

REFERENCES

Cicchitelli G., Herzel A., Montanari G. E., (1992) , Il campionamento statistico, Il Mulino.

Fay R. E., (1996), Alternative paradigms for the analysis of imputed survey data, Journal of the American Statistical Association, 91,426,pp.490-498.

Ford B. L., (1983), An overview of hot-deck procedures, in: Incomplete data in sample surveys, Madow W. G., Olkin I., Rubin D. B.(Eds.) , Academic Press, New York, pp.185-207.

Kalton G. and Kasprzyk D.,(1986),“The treatment of missing data ”,Survey Methodology, 12, pp.1-16.

Little R. J. A. and Rubin D. B., (1987), Statistical analysis with missing data, JohnWiley , New York .

Monte A. and Schoier G., (1999), Metodi d’imputazione delle non risposte: confronti e verifiche, Convegno della Società Italiana di Statistica: Verso i Censimenti del 2000, Udine, 1999,(abstract).

Rao J. N. K. and Shao J., (1992), Jackknife variance estimation with survey data under hot deck imputation, Biometrika, 79,4,pp.811-822.

Rao J. N. K., (1996), On variance estimation with imputed survey data Journal of the American Statistical Association, 91,434,pp.499-506.

Rubin J. N. K., (1978), Multiple imputations in sample surveys: a phenomenological bayesianapproach to non response , in Procedings of the Survey Research methods Section, American Statistical Association, ,pp.20-34.