

Auxiliary Variables in Income Imputation by Minimum Distance Donor Method⁽¹⁾

*L'Informazione Ausiliaria nell'Imputazione del Reddito. Il Metodo del
Donatore con Distanza Minima*

Claudio Quintano, Rosalia Castellano, Andrea Regoli
Istituto di Statistica e Matematica, Università di Napoli "Parthenope"

Via Medina, 40 - 80133 Napoli

claudio.quintano@uninav.it;

lia.castellano@uninav.it;

andrea.regoli@uninav.it

Riassunto: In questo lavoro si intende valutare il processo di ricostruzione della variabile reddito familiare a seguito dell'imputazione per la presenza di mancate risposte, evidenziando il contributo delle variabili ausiliarie nei metodi proposti. A tale scopo, si restringe l'attenzione ai metodi hot-deck, che sono frequentemente utilizzati dagli Istituti di statistica ufficiali. In particolare, si considera l'imputazione da donatore con funzione di distanza mista, la cui performance viene posta a confronto con quella realizzata con il metodo da donatore casuale all'interno di classi di imputazione. La strategia adottata prevede la simulazione di dati di reddito mancanti, a partire dalle informazioni dell'Indagine sui Bilanci delle Famiglie Italiane della Banca d'Italia relativa al 1998.

Keywords: income data, nonresponse mechanism, hot-deck methods.

1. Introduction

In this work we face the problem of item nonresponse affecting income data. Our aim is to evaluate the performance of an hot-deck imputation method based on a distance function matching (minimum distance donor method) compared with a random hot-deck within classes method. We concentrate on hot-deck methods since they are widely used in their different specifications by National Statistical Institutes.

The accuracy of the reconstruction process is measured both in terms of microdata and statistics of the whole distribution. The strategy we follow is based on simulation techniques. The starting point of the process is the perturbation of a real dataset where some values are made missing through an *ad hoc* mechanism.

2. Simulation of missing income data

On the basis of the overall sample (7,147 units) of the *Survey of Italian Household Income and Wealth* in 1998, *item missing data* are artificially created in the annual

⁽¹⁾ This paper is financially supported by MURST 60%, year 1999, "Control and correction processes of economic data", Coordinator: prof.ssa Rosalia Castellano, Università degli studi di Napoli "Parthenope".

household *income*. This survey provides a rich basis of information, useful as auxiliary socio-economic tool for hot-deck methods (Ford, 1983).

Nonresponse rate is 20 percent. The introduced MAR mechanism relates the income response probability and the consumption, as the Bravais-Pearson linear correlation coefficient shows the strongest value between income and consumption ($r=0,69$).

Nonresponse probability p_i is specified through the following function:

$$p_i = 1 - e^{-I\left(\frac{x-Q_1}{s}\right)} \quad \text{if } x > Q_1 \quad \text{or} \quad p_i = 0 \quad \text{if } x \leq Q_1$$

where x is the consumption, $Q_1=21,400$ is the first quartile and $\sigma=26,182$ is the standard deviation of its distribution, $\tilde{e}=0.443$ is chosen in order to get 20% of missing data.

Through the missing data generating process, the original distribution is divided into two subsets, the respondents distribution ($n_R=5,691$) and the artificial nonrespondents distribution ($n_M=1,421$). The number of repeated simulations is $s=10$, and this seems a reasonable convergence threshold for the stability of the results.

3. Hot-deck imputation

The hot-deck methods based on a distance function, as developed in RIDA procedure (*Ricostruzione delle Informazioni con Donazione Automatica*; Abbate, 1997), evaluate the similarity between units through a metric that depends on the measurement scale of each variable. The set of variables can be divided in stratification and matching variables. The former define groups of homogeneous observations so that the donor's search is restricted within these groups. The latter are the variables which actually enter in the mixed distance function. Within each stratum, an item missing value is replaced by the value observed for the unit which has the smallest distance from the incomplete unit among all the potential donors.

In order to stress the role of auxiliary variables in the reconstruction process through the distance hot-deck method, the random hot-deck method within imputation classes is chosen as a benchmark. The random hot-deck method stratifies the units according to the observed variables: strata are then defined through a combination of categories and the donor is to be searched within the same stratum where the incomplete unit is classified.

The sub-sample of respondents to income variable (5,691 units) is the donors' set. In order to apply the distance function hot-deck method, RIDA software has been used, which is implemented in CONCORD (*CONtrollo e CORrezione Dati*) system, developed in SAS environment at ISTAT. A first, very general specification (*DHD1, Distance Hot-Deck 1*) includes all the variables that satisfactorily correlate with income; they all are treated as matching variables and they get the same weight in the definition of the mixed distance.

The performance of this specification can be compared with a refined version of the method, where a selection of available information is made and a stratification variable is introduced (*DHD2, Distance Hot-Deck 2*). The matching variables reduce to four: household consumption, household wealth, education and work status of head of household (hh). They have been chosen because of their strongest correlations with income. Every matching variable gets a weight proportional to the measure of its

association with income. As stratification criterion, the number of earners in the household is considered.

Random hot-deck methods within classes have been implemented through SOLAS software. The first specification (*RHD1, Random Hot-Deck 1*) refers to a classification based on the same variables entered in the distance function method, that is number of earners, household consumption, household wealth, education of hh and work status of hh. In order to evaluate the effect of the grouping in consumption classes (the variable which the nonresponse mechanism depends upon), a second method is specified (*RHD2, Random Hot-Deck 2*), that differs from the previous one only with regard to the treatment of consumption expenditure, now grouped into 3 instead of 5 classes. In the end, due to the sensitivity of random hot-deck within classes method to the hierarchy in stratification variables, a third specification (*RHD3, Random Hot-Deck 3*) is introduced, which gives priority to the head of household characteristics.

4. Evaluation criteria and results

The different imputation methods can be compared as to the accuracy in the reconstruction of: i) the main descriptive statistics of the whole income distribution and ii) the association measures with income. The evaluation criteria are the relative bias of statistics such as position indexes, mean, standard deviation and skewness. The relative bias is also evaluated of the correlation measures between income and the main auxiliary variables. In the overall distribution, the bias in a given statistic T is $Bias(T) = E(T_{rec,s}) - T_{orig}$, measured as the difference between the recovered value T_{rec} , averaged across $s=10$ replications, and the corresponding original value T_{orig} .

A further measure of how well an imputation method performs, with special reference to the preservation of single values after the imputation, is the average imputation error (*AIE*), defined as the root mean squared deviation (*RMSD*) between imputed and real income values, averaged across $s=10$ replications:

$$AIE = E(RMSD_s) = E \left(\sqrt{\frac{1}{n_M} \sum_{i=1}^{n_M} (Y_{i,imp,s} - Y_{i,real,s})^2} \right)$$

As to the preservation of income distribution, all the statistics of the reconstructed distribution underestimate the corresponding original indexes; no matter what imputation method is used, the relative biases have always negative signs (see Table 1). By comparing the two methods based on the distance function, the bias is always smaller when we apply the more refined procedure (*DHD2*). This method also shows better results in the mean, variability and skewness indicators than the three random hot-deck specifications.

In detail, the random hot-deck methods produce a strong reduction in the original variability, and *RHD1* has the worst performance (-18.78%); the choice of consumption as auxiliary stratification variable (grouped into 3 or 5 classes), together with the dependence of nonresponse mechanism on consumption level, has the effect of letting the number of donors very small in the highest consumption classes. Consequently, in those classes there is a high probability of using repeatedly the same records as donors, and this causes a decrease in the variability.

Table 1: *Relative biases in parameter estimates and average imputation error*

	Imputation methods				
	DHD1	DHD2	RHD1	RHD2	RHD3
Relative biases in:					
-1st quartile	-2.13	-0.29	-0.27	-0.24	-0.57
-Median	-3.49	-0.51	-0.39	-0.79	-0.73
-3rd quartile	-5.06	-0.84	-0.82	-0.74	-0.84
-9th decile	-6.57	-2.55	-2.34	-2.21	-2.19
-Mean	-7.03	-3.37	-3.63	-3.55	-3.63
-Standard deviation	-23.53	-16.25	-18.78	-17.85	-18.42
-Skewness	-51.01	-35.19	-43.69	-41.09	-41.95
-Correlations:					
income-consumption	-27.77	-17.18	-22.66	-23.60	-23.88
income-wealth	-22.30	-12.10	-18.31	-19.23	-21.06
income-education of hh	8.08	4.82	6.35	5.72	6.16
income-work status of hh	3.64	1.43	2.19	2.27	3.19
Average imputation error	63,437	61,214	64,529	65,134	65,721

In order to analyse the effects on the correlations, we note that all the imputation methods attenuate the association measures with consumption and wealth, even if these variables enter as auxiliary variables in the method's specification. Among all methods, *DHD2* method has the smallest bias in the correlations with consumption and wealth. The bias in the associations with education and work status is always positive and *DHD2* method still shows the lowest values. Even when referring to the preservation of single values, *DHD2* method is still the preferred one (see *AIE* values).

The *RHD2* and *RHD3* specifications, that differ only in the order of the auxiliary variables as stratification variables, give the same performance indicators.

Extremely difficult for every method is the recovery of high income values: the bias in the ninth decile is very large, compared with the other percentiles. By further disaggregating high income households on the basis of auxiliary variables, some groups in the population have been identified who are at risk of heavy underestimation should their income not be observed, but imputed. Such groups include households whose head is graduated and works as manager or sole proprietor or member of arts or professions; for the imputation of their income, one can envisage alternative treatment strategies, resulting from a mixture of automatic and interactive methods, and allowing the researcher's intervention to guide the correction system towards the finding of the most "critical" strata.

References

- Abbate C. (1997), La completezza delle informazioni e l'imputazione da donatore con distanza mista minima, *Quaderni di ricerca*, 4, 67-102, Istat, Roma.
- 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.), vol. 2, 185-207, Academic Press.