

# **Multivariate Methods for Item Nonresponse: an Application to the Italian Household Budget Survey**

*Metodi Multivariati per il Trattamento delle Mancate Risposte Parziali:  
una Applicazione all'Indagine Istat sui Consumi delle Famiglie*

Claudio Ceccarelli  
ISTAT  
V. A. Ravà 150  
00142 Roma  
e-mail: clceccar@istat.it

Fabio Massimo Rottino  
ISTAT  
V.le Liegi 13  
00100 Roma  
e-mail: rottino@istat.it

**Riassunto:** Il problema delle mancate risposte parziali è stato affrontato attraverso l'imputazione da modello; in particolare, un metodo di imputazione esplicita con il modello di regressione lineare è stato sperimentato sui dati dell'indagine ISTAT sui consumi delle famiglie italiane relativi al 1997. Il confronto dei valori imputati con quelli originalmente presenti ha permesso di valutare la bontà del metodo proposto.

**Keywords:** Item nonresponse, Regression imputation, Missing data, Household budget survey

## **1. Introduction**

Nonresponse constitute one of the principal causes of nonsampling error (Cicchitelli, Herzal, Montanari, 1992). The information and nature of nonresponse, in some countries, are an integral part of the quality system of survey institute. Probably, this will be the future trend (Laaksonen, 1996).

There are two principal standard techniques to face the nonresponse problem: reweighting and imputation (Särndal, 1996). The model-based approach for imputing missing values takes part in so-called explicit imputation methods, that is all those techniques that use the principle of finding a suitable substitute for each missing value (Little, 1982). So, common objective for all explicit imputation methods is to search a substitute value the more similar to the missing one. The definition of assigning variables, statistically linked with that who present missing value, grants this likelihood.

## **2. The imputation model by linear regression**

The linear regression model for the imputation of missing values can be regarded as a particular case of the more generalized form outlined by Kalton and Kasprzyk (1982), that links auxiliary data (observed for all sampled units) with the item's missing value. In its more general form, it can be wrote in this way (Lessler and Kalsbeek, 1992):

$$Y_i = \hat{\beta}_0 + \sum_{j=1}^K \hat{\beta}_j X_{ji} + e_i$$

where  $j=1,2,\dots,k$  are the assigning variables.

Suppose that  $Y_1, Y_2, \dots, Y_k$  are  $K$  variables multivariate normally distributed with mean  $\mu$  and covariance matrix  $\Sigma$ ; then the missing variables can have linear regression on the observed variables, with regression coefficients that are well-known functions of  $\mu$  and  $\Sigma$ . The method proposed by Buck (1960) first estimate  $\mu$  and  $\Sigma$  from the sample mean and covariance matrix based on the complete cases, and then uses these estimates to calculate the linear regression of the missing variables on the present variables, case by case (Little and Rubin, 1987).

If the values are missing completely at random, that is in case of ignorable missing-data mechanism (Rubin, 1974, 1976; Little, 1983), the Buck's method provide a consistent estimates of the averages of observed and imputed values. The sample covariance matrix from the filled-in data, instead, underestimates the size of variances and covariances; in fact, it sets to zero the residual variance for the filled-in values, since the imputations lie exactly on the regression line.

### 3. The experimentation

The Buck's imputation model has been experimented on data of the Italian household budget survey relative to year 1997 (Ceccarelli *et al.*, 2000). The imputation of the missing value has not been performed by substitution with the value of another statistical unit (nearest neighbour), such as behave, in extreme synthesis, the RIDA algorithm, actually used by ISTAT in this survey (Kalton and Kish, 1981). In our experimentation, a linear regression model has been constructed on complete observation; the dependent variable will be that subject to item nonresponse for some observation. Parameters obtained are so used as dependent variables for constructing an explicit linear imputation model, in function of those values for whom we have information and that play the role of independent variable. The missing values have been artificially produced, being careful to save the original value to allow a subsequent face up to the estimates (Quintano e Lucev, 1990). The complete data matrix has been subjected to a random erasure of a certain expense item chosen for 1.000 families on a total of 22.362, equal to 4,5% rate of nonresponse. The model's strength has been tested in the following three ways:

1. it has been supposed that missing values spread out in five different distribution functions: Uniform, Normal, Poisson, Cauchy and Exponential;
2. 30 simulations have been experienced for each one of the previous distribution functions;
3. points 1. and 2. have been applied to three different expense items.

Three expense items have been chosen to impute the missing values artificially produced: CAP1 (bread and cereals), CAP11 (house) and CAP12 (combustible and electricity). These items, in fact, are less subject to the expense's zeroes and, at the same time, present different characteristics of variance and values of budget's averages.

Clusters have been formed according to the budget's values and a stepwise regression has been used in each one of the clusters; expense's profiles, that is the expense for each

item of each family divided by the total expense of that family, have been utilized in place of the real values,. So, in every cluster the expense behaviour can be assumed nearly homogeneous; the parameters of the regression linear model on the expense item of the complete observation in a specific cluster are the most suitable for constructing an explicit linear imputation model for those families that have a missing value in an expense item but, according to the values of the other items, have been enclosed exactly to that cluster. The last step is to obtain the missing value, i.e. of the first expense item for the  $i$ -th family in the  $c$  cluster, according to the following model:

$$\hat{Y}_{i,1,c} = \hat{\beta}_{0,c} + \sum_{j=2}^{19} \hat{\beta}_{j,c} X_{i,j,c} ,$$

$\hat{\beta}_{0,c}$  is the intercept and  $\hat{\beta}_{j,c}$  the parameter of the  $j$ -th expense item in the  $c$  cluster;  $X_{i,j,c}$ , instead, is the value of the  $j$ -th expense item for the  $i$ -th family in the  $c$  cluster.

The same procedure has been used to obtain parameters within the 11 groups of familiar typology utilized by ISTAT for grouping Italian families, and find the following second estimator for the  $i$ -th family in the  $g$  group:

$$\hat{Z}_{i,1,g} = \hat{\beta}_{0,g} + \sum_{j=2}^{19} \hat{\beta}_{j,g} X_{i,j,g} .$$

The third estimator, finally, has been achieved as an average of the complete estimates' vector within all the 11 groups of familiar typology, weighted with the respective values of the probability that  $i$ -th family belong to the different groups. These probability values have been gained by a discriminant analysis:

$$\hat{W}_{i,1} = \sum_{g=1}^{11} \hat{Z}_{i,1,g} p_{i,g} .$$

A distance relative index (IDR) has been performed to match the estimates with the original values; the IDR is the ratio of the absolute difference between value's estimate and original value, to the absolute value of the maximum of both values (Barcaroli *et al.*, 1998).

#### 4. Results and conclusions

The averages on the 1.000 families of the imputed values by means of this model satisfactorily approach the averages of the true values on the same families; and this happen for each expense item and for every kind of distribution of the missing values. The IDR, in fact, has a very restricted range [0,97%; 3,56%], with a mean rate of 2%. The best fitness is obtained for the expense relative to bread and cereals, with both lower average monthly expense and variance. This item shows, in addition, an excellent matching achieved for the Normal distribution and the estimators  $\hat{Y}_1$  and  $\hat{Z}_1$ .

The relatively worst fitness is performed by the house expenses, involving a higher budget level and a wider variance than the other items. The Normal distribution always fits better than the others, but now the best results have been produced with the weighted estimator  $\hat{W}_{11}$ .

The results obtained in the case of the expense for combustible and electricity lay in an intermediate position among the previous expense items.

## References

- Barcaroli G., Luzi O., Ceccarelli C. (1998). Il macroediting: tecniche guidate di correzione interattiva di variabili quantitative guidata dall'analisi degli aggregati. Il caso del sistema dei conti delle imprese, *Quaderni di ricerca*, ISTAT, n.1/1998.
- Buck S.F. (1960). A Method of Estimation of Missing Values in Multivariate Data Suitable for Use with an Electronic Computer, *Journal of the Royal Statistical Society*, B 22, pp. 302-307.
- Ceccarelli C., Coccia G., Grassi D., Pannuzi N. (2000). Il Trattamento dei Dati nel Processo di Produzione dell'Indagine sui Consumi delle Famiglie, di prossima pubblicazione sulla rivista *Scritti di Statistica Economica* Vol. 7.
- Cicchitelli G., Herzel A., Montanari G.E. (1992). *Il campionamento statistico*, Il Mulino, Bologna.
- Kalton G. e Kish L. (1981). Two efficient random imputation procedures, *American Statistical Association*, Proceedings of the Survey Research Methods Section, pp. 146-151.
- Kalton G., Kasprzyk D. (1982). Imputing for Missing Survey Response, *American Statistical Association*, Proceedings of the Survey Research Methods Section, pp. 22-31.
- Laaksonen S. (1996). Preface: Nonresponse – An Essential Indicator of Survey Quality, in *International Perspectives on Nonresponse*, Proceedings of the Sixth International Workshop on Household Survey Nonresponse, Statistics Finland, Helsinki.
- Lessler J.T., Kalsbeek W.D. (1992). *Nonsampling Error in Surveys*, Wiley, New York.
- Little R.J.A. (1982). Models for Nonresponse in Sample Surveys, *Journal of the American Statistical Association*, V. 77, pp. 237-250.
- Little R.J.A. (1983). The Ignorable Case, The Nonignorable Case, in *Incomplete Data in Sample Surveys*, Vol. II: Theory and Annotated Bibliography (W.G. Madow, I. Olkin, D.B. Rubin, Eds.), Capitoli 21-22, Academic Press, New York.
- Little R.J.A., Rubin D.B. (1987). *Statistical Analysis with Missing Data*, Wiley, New York.
- Quintano C., Lucev D. (1990). Le Mancate Risposte in Esperienze di Indagini Redditali, *Quaderni sardi di Economia*, n° 3.
- Rubin D.B. (1974). Characterizing the Estimation of Parameters in Incomplete Data Problems, *Journal of the American Statistical Association*, V. 77, pp. 467-474.
- Rubin D.B. (1976). Inference and Missing Data, *Biometrika*, 63, pp. 581-592.
- Särndal C.E. (1996). For a Better Understanding of Imputation, in *International Perspectives on Nonresponse*, Proceedings of the Sixth International Workshop on Household Survey Nonresponse, Statistics Finland, Helsinki.