

Proper Imputation of Missing Income Data for the Tuscany Living Condition Survey

Imputazione Propria per Dati di Reddito Mancanti nell'Indagine sulle Condizioni di Vita in Toscana

Giulio Ghellini

Università degli Studi di Siena, Piazza S. Francesco 8, Siena, ghellini@unisi.it

Laura Neri

Università degli Studi di Siena, Piazza S. Francesco 8, Siena, neri@unisi.it

Riassunto: Questo lavoro riguarda due aspetti fondamentali dell'Indagine sulle Condizioni di Vita in Toscana (ICVT): il metodo di rilevazione dei dati di reddito e l'imputazione dei dati di reddito da lavoro mancanti. Per quanto riguarda il primo aspetto, si è optato per una tecnica di rilevazione mista (redditi lordi provenienti dalla documentazione fiscale oppure redditi netti autodichiarati) determinata in base all'esito dell'indagine pilota in cui erano state sperimentate le due diverse modalità di rilevazione. Per quanto concerne il trattamento dei dati mancanti, si è optato per tecniche di imputazione differenziate a seconda della situazione reddituale dichiarata dagli intervistati. Un semplice esercizio di validazione della procedura di imputazione adottata conclude il lavoro.

Keywords: income data collection, nonresponse, imputation.

1. Introduction

Sampling statistician and survey researchers agree that nonresponse in sample survey is a source of serious errors. Biases in point estimators, mainly due to the difference between respondents and nonrespondents, receives the principal attention in survey literature; however inflation of variance in point estimators and biases in estimators of precision are also of importance in performing inference procedures. In this paper we deal with data from the sample survey on Income and Living Conditions (ICVT) conducted on the Tuscany households, in particular the missing data treatment. A peculiarity of the survey ICVT, presented in Section 2, is the collection of incomes using “if possible” individual taxation form. Section 3 presents the missing data treatment adopted for income missing data and some suggestion for further research work.

2. The income data collection

Questions about personal or household income are notorious for eliciting refusals and also for “not careful” answers. Considering these problems, it has been decided to test two different way of collecting such information by a preliminary pilot survey:

- a) asking to the individual the *net labour income*, following the Bank of Italy scheme;
- b) asking to the individual the *gross labour income* taken from the taxation forms.

The analysis of the pilot survey data and the debriefing with the interviewers showed that the (b) way provides more careful and reliable answers even if it creates some problems with elder people (frequently forgetting of having taxation form) and with people having not regular work; instead, the (a) way gave a higher response rate but with clear evidence of under-reporting.

For these reasons it has been decided to adopt a mixed strategy. For retired people, for individuals having only wage or salary earnings and for not regular workers (not obliged to compile taxation form), the questionnaire gave the possibility of reporting the net income instead of the gross income indicated in their official taxation forms. For all the other individuals having more complex fiscal situations the questionnaire asks gross income referring to their specific taxation form (CUD, 730, UNICO).

3. Imputation of the missing income data

In most current surveys, nonrespondent adjustment assumes that the missing data mechanism is Missing At Random –MAR– (Rubin, 1976); this strong assumption is usual when the treatment of nonresponses regards public user data bases (see Rubin, 1987 for general references, and Quintano *et al.*, 2001 for application on Italian income data); for this reason also in this work it is assumed the MAR assumption. Before introducing and explaining the imputation procedures applied, it is crucial to consider the response behaviour to the question (C3.2) asking the Taxation Form compiled by the interviewed. People reporting the net income perceived presented lower nonresponse rate (on the average the 11%) than individuals indicating the gross income reported on the official taxes form (CUD, 730, UNICO), in fact in this case the average nonresponse rate is about 41%. The response behaviour according to the classification of the question C3.2 is different either for the amount of nonrespondents either for the average income reported (see Ghellini, Neri 2004 for details). In order to test the significance of the explicative power of the variable C3.2 an Analysis of Variance has been conducted on the observed individual labour income considering C3.2 as classification variable. The analysis has shown that about the 10% of the income variability is explained by the classification variable. These results suggest to perform the imputation procedure as follows:

- for individuals belonging to groups that gave net income (groups 1, 2, 6) it has been performed a conditional mean imputation, for each of the three groups, by cell defined by individual sex (male and female) and educational level (low and high);
- individuals belonging to the groups reporting the gross income (groups 3, 4, 5) are characterized by higher nonresponse rate, with respect to the previous group; moreover average income and variability of the observed income are different among groups. For this reason each subgroup has been dealt independently on the others. The individual labour income has been transformed in log terms in order to make more reasonable the usual normality assumption. The imputation has been performed applying a proper stochastic regression imputation method (Rubin, 1987), in which each missing values is replaced by its regression prediction plus a

random error with variance equal to the estimated residual variance. The results of the regression applied to the observed data for each fiscal model are in Table 1.

*Tab.1: Regression models for income according to each taxation Form: parameter estimates and significance levels (*** p -value ≤ 0.01 ; ** $0.01 < p$ -value ≤ 0.05)*

Taxation Form	CUD (3)	730 (4)	UNICO (5)
Variable ¹	Estimate (p.v)	Estimate (p.v)	Estimate (p.v)
Costant	15.06891 ***	15.09531 ***	14.48347 ***
Dsesso (1=male)	0.35609 ***	0.26180 ***	0.29490 ***
etax1 (age)	0.04108 ***	0.05662 ***	0.06604 ***
etax1sq (age ²)	-0.00032 **	-0.00049 ***	-0.00053 ***
dstu23 (1=middle school)	0.41098 ***	0.27282 ***	0.28040 ***
dstu4 (1=secondary school)	0.56448 ***	0.48468 ***	0.55356 ***
dstu56 (1=university degree)	0.72220 ***	0.65234 ***	0.66640 ***
dc3101 (1=individual declaration)		-0.18906 ***	
dposdip (1= in paid employment)		0.20760 ***	0.21543 **
Dposind (1=self employment)			0.22164 **
<i>Adj R-Sq</i>	<i>0.1525</i>	<i>0.3216</i>	<i>0.2567</i>

Estimated average income and related variability measure computed on observed and completed data are in Table 2. It can be observed that completed values generally present, as expected, higher average income; regarding the variability it decreases when a deterministic imputation procedure is applied but it generally increases by applying a stochastic imputation procedure.

*Tab.2: Individual net labour income -yearly income in thousands of lira- on observed and completed data (*as observed data are considered previously imputed data too)*

Taxation Form (C.3.2)	Observed Data				Completed Data			
	N	Mean	Std	v.c.	N	Mean	std	v.c.
1=None, only retirement income	787	15468	10222	0.661	817	15526	10072	0.649
2=None, only wage	245	20428	8683	0.425	264	20478	8389	0.409
3=CUD 2001	254	22893	11663	0.509	370	23024	13369	0.581
4=730/2001	1057	25000	13392	0.536	1654	26145	14626	0.559
5=UNICO 2001	499	29954	35143	1.173	1068	32800	31881	0.972
6=No taxation form at all	74	18217	17083	0.938	174	17720	12062	0.681
‘Missing’*	4347	24837	20277	0.816	4703	24885	20249	0.814

As a validation of the imputation procedures applied, the membership function to the fuzzy set of poor has been computed, with respect to the so called ‘supplementary variables’ (Betti *et al.*, 2000). The analysis has been conducted separately for the set of poor and non-poor, defined according to the household equivalent income (traditional approach). Results in Table 3 shows that on the averages people having imputed income and belonging to the set of poor are characterized by a membership function to the

¹ The available data do not allow more detailed classification of the individual work position than “in paid employment” *versus* “self employment”.

fuzzy set of poor very similar to the observed one. The same happened for the non-poor. It has to be noticed that the single imputation procedure applied, suffers of a serious defect: the single imputed value cannot represent all of the uncertainty about each value to impute, for this reason, it underestimates uncertainty even if nonresponse is modelled correctly and random imputations are created. Large samples results (Rubin, Shenker, 1986) show that for simple situation with 30% of the missing data, single imputation under the correct model results in nominal 90% confidence intervals having actual coverage below 80%. A next step of this work will be to take into account of this deficiency considering two approaches. The former consists of applying a replication method of variance estimation and recompute the imputations on each replicate sample (Rao, 1996), the latter is the multiple imputation procedure (Rubin 1987).

Tab.3: Membership function to the fuzzy set of poor with respect to the supplementary variables

	Mean	Median	V.C.
<i>Poor Set (traditional approach)</i>			
Completed	0.233	0.198	0.875
Observed	0.239	0.199	0.858
Imputed	0.217	0.190	0.918
<i>Non Poor Set (traditional approach)</i>			
Completed	0.153	0.129	0.949
Observed	0.155	0.130	0.931
Imputed	0.150	0.125	0.969

References

- Betti G., D'Agostino A., Neri L. (2000) Panel Regression Models for Measuring Poverty Dynamics in Great Britain, ISER Working Papers 2000-42, University of Essex.
- Ghellini G., Neri L. (2004) Indagine sulle Condizioni di Vita delle Famiglie Toscane: il disegno d'indagine ed il processo di produzione dei dati, *mimeo*.
- Quintano C., Castellano R., Regoli A. (2001) How to improve the quality of the income variable in a household survey. A simulation study through multiple imputation, *Proceedings of the International Conference on Quality in Official Statistics*, May 2001, Stockholm (CD-ROM).
- Rao J. N. K. 1996 On Variance Estimation with Imputed Survey Data, *Journal of the American Statistical Association*, 91, 499-506.
- Rubin D. B. (1976) Inference and Missing Data, *Biometrika*, 63, 581-590.
- Rubin D. B. (1987) *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley.
- Rubin D. B., Shenker N. (1986) Multiple Imputation for Interval estimation from Simple Random Samples with Ignorable Nonresponse, *Journal of the American Statistical Association*, 81, 366-374.