An evaluation of unit nonresponse bias in the Italian Households Budget Survey

Ceccarelli, Claudio; Coccia, Giuliana; Crescenzi, Fabio

Veröffentlichungsversion / Published Version
Konferenzbeitrag / conference paper

An Evaluation of Unit Nonresponse Bias in the Italian Households Budget Survey

CLAUDIO CECCARELLI, GIULIANA COCCIA AND FABIO CRESCENZI

Abstract: *The effect of nonresponse is a crucial aspect which has received considerable attention in literature. Nonresponse causes an increase in sampling variance and the estimates of units nonresponse bias are useful to give a concrete measure of this increase. Scarce co-operation of respondents gives inaccurate or incomplete information, and this part of the answers can't be used in the estimation. Since 1991, Istat erases from its sample file all households with incompatible data on expenditure and income. We call this the "UR (Unreliable Respondents) sub-population". UR data can't be used to build estimates, but can be used to give approximate information on NR. The idea is that, concerning behaviour, the NR population is more similar to UR than to R.*
The results of an attempt to evaluate unit nonresponse bias in the 1995 Italian Household Budget Survey (HBS) are shown in this paper. These results were obtained without performing a specific survey on nonrespondents.

Keywords: *unreliable respondents, unit nonresponse bias*

1 Introduction

The results of an attempt to evaluate the unit nonresponse bias in the 1995 Italian Household Budget Survey (HBS) are shown in this paper. These results were obtained without performing a specific survey on nonrespondents.

Quality of estimates depend on sampling and non sampling errors and one of the major components affecting non sampling errors is the rate of non response (Platek and Gray 1986). A population can be divided in respondents (R) and nonrespondents (NR), but there are several difficulties in the correct evaluation of non response bias because of the lack of information on NR. The high cost of additional surveys force statisticians to look for new methods to build approximate estimates of bias (Lessler and Kalsbeek 1992).

Scarce co-operation of respondents gives inaccurate or incomplete information, and this part of the answers can't be used in the estimation. Since 1991, Istat erases from its sample file all households with incompatible data on expenditure and income. We call this the "UR (Unreliable Respondents) sub-population". UR data can't be used to build estimates, but can be used to give approximate information on NR. The idea is that,

concerning behaviour, the NR population is more similar to UR than to R.

Non respondent bias is affected by two components. We call V_I or Variable Independent, the first of these two components, which is independent from the value of the considered variable in the NR population. The second one, which we call V_D or Variable Dependent, depends on the differences between values assumed by the considered variable in the R and in the NR populations.

Concerning the first component, NR data are usually available, concerning the second one, it must be estimated and our approach is to use UR data to approximate NR data.

2 HBS survey design

HBS gives a continuous flow of data on Italian households' expenditures, incomes and other characteristics (Istat 1995). There are two main survey techniques on which data collection is based:

- a self-compiled diary. For ten consecutive days consumers record detailed information on expenditures, aiming especially at small purchases, which are often difficult to recall;
- a face to face interview that, at the end of the month, allows information on large expenditures to be collected. It also provides data on family size, dwelling characteristics, durable consumer goods, etc.

The sample scheme is a two stages stratified sample. Communes represent the Primary Sample Units (PSUs). Every quarter are 550 PSUs (out of 8103 Italian Communes) surveyed, stratified by demographic size, altitude zone and prevalent economic activity.

The PSUs of each geographical region are divided in two groups:

1. Communes with more than 50,000 inhabitants, permanently included in the sample;
2. other Communes, included in the sample on a rotating panel basis (one month quarterly).

PSUs are selected from each stratum without replacement and probabilities of inclusion proportional to size. Every quarter 148 PSUs come from the first group and 402 PSUs from the second one.

Secondary Sample Units (SSUs) are households listed in municipal registers. SSUs are sampled using a systematic scheme, without replacement and with equal probabilities of inclusion. All the components of selected households are included in the sample. Every

quarter about 9,000 households are interviewed (yearly about 36,000 households).

To ensure obtaining the fixed sample size, nonrespondent households are replaced by other households of the same PSU.

Quarterly expenditure estimates are available at the national level, while annual estimates are available at the regional level.

The estimator used to produce each of the estimates is post-stratified by household size.

3 HBS response rates and unreliable respondents

Table 1 shows the percentage of households, response rates and mean expenditures by household size.

Table 1: 1995 HBS-percentage of households, response rates and mean expenditure, by household size

	Household Size				
	One component	Two components	Three components	Four or more components	Overall
Households (%)	25.6	29.2	22.3	22.9	100.0
Response rates (%)	86.8	91.8	93.6	94.6	91.3
Mean expenditure (thousands of liras)	1,886	2,763	3,687	4,078	3,2218

It can be highlighted that, not surprisingly, the response rate is an increasing function of the number of components: one component households reach the lowest rate, four or more component households attain the highest.

Nonrespondents can be divided by nonresponse categories. Table 2 shows the percentage of households by those categories.

Table 2: 1995 HBS-percentage of NR households by nonresponse categories

	Percentage
Cannot be found	26.3
No one at home	23.3
Refusals	32.6
Others	17.8
Total nonrespondent	100.0

The "Cannot be found" category includes those households which couldn't be contacted for the entire survey period (i.e. wrong address). The "No one at home" category includes households which couldn't be found at home after repeated attempts at various times. The "Refusals" category includes households refusing to participate in the survey.

Table 3 gives the distribution of nonrespondents by nonresponse categories and household size.

Table 3: 1995 HBS-percentage of NR households by nonresponse categories and household size

Household Size	Nonresponse categories				
	Cannot be found	No one at home	Refusals	Others	Total Nonrespondents
One person	40.0	35.3	23.4	46.1	33.8
Two persons	22.6	26.7	28.5	23.1	25.7
Three persons	18.2	16.2	22.9	13.4	18.7
Four or more persons	18.6	21.5	24.8	15.1	21.0
Not indicated	0.6	0.3	0.4	2.3	0.8
Total	100.0	100.0	100.0	100.0	100.0

Identification of unreliable households (UR) is based on the comparison of data on food and non-food expenditures, income and family size.

Households showing very low food and non-food expenditures, and a high percentage of the non response item is labelled as unreliable.

Unreliable data are deleted from the original file and are not used to estimate expenditures.

Dividing population in two sub-sets: r = RR and $\bar{r} = (NR + UR)$, it is possible to define the nonresponse rate as:

$$T = \frac{\bar{r}}{r + \bar{r}} \times 100 = \frac{NR + UR}{RR + NR + UR} \times 100 \qquad (1)$$

Table 4 shows the number of UR, NR, R and T by region. T assumes the highest values in Friuli V.G. (17.7%) and Piemonte (13.9%), the lowest values in Marche (4.8%) and Puglia (6.8%).

Table 4: 1995 HBS-nonresponse rates by region

Region	Respon-dents RR	Nonrespon-dents NR	Unreli-able UR	$T = \frac{NR + UR}{NR + UR + RR} \times 100$
Piemonte	2,631	382	44	13.9
Valle d'Aosta	936	41	35	7.5
Lombardia	3,211	468	40	13.7
Trentino Alto Adige	1,940	208	27	10.8
Veneto	1,760	117	26	7.5
Friuli Venezia Giulia	998	199	16	17.7
Liguria	1,713	168	12	9.5
Emilia Romagna	2,029	258	43	12.9
Toscana	2,437	232	35	9.9
Umbria	1,035	58	32	8.0
Marche	1,416	41	30	4.8
Lazio	2,418	175	69	9.2
Abruzzo	960	38	57	9.0
Molise	1,028	77	32	9.6
Campania	2,564	116	98	7.7
Puglia	1,924	74	67	6.8
Basilicata	923	15	73	8.7
Calabria	1,251	118	41	11.3
Sicilia	2,334	146	110	9.9
Sardegna	895	71	41	11.1
ITALY	34,403	3,002	928	10.3

4 Evaluation of unit nonresponse bias

1995 HBS expenditure estimates are based on a family size post-stratified estimator and a two stage stratified sample. Post stratification reduces the effects of nonresponse on estimates (Cochran 1977; Holt and Smith 1979).

In this paper, we show the results of an evaluation of bias obtained by using a simplified sample scheme, which does not consider the first stage PSUs, but only a stratification by regions of sampled households and a post stratification by household size. This is to reduce the formula's complexity, with slight impact upon results. However, it is easy to generalise formulas to the more complex case.

Let us call:

h	stratum index (regions) (h=1,...,H);
j	SSUs (households) by stratum index (j=1,..., M_h);
l	size of households class (l=1,...,L);
M_{hl}	number of SSUs of size l in the stratum h;
m_{hl}	number of sampled SSUs of size l in the stratum h;
M_h	number of SSUs in the stratum h;
m_h	number of sampled SSUs in the stratum h;
M	number of SSUs;
m	number of sampled SSUs;
y_{hlj}	expenditure y of the j SSU of size l in the stratum h.

The mean population value of y is then:

$$\bar{y} = \frac{1}{M} \sum_{h=1}^{H} \sum_{l=1}^{L} \sum_{j=1}^{M_{hl}} y_{hlj} \tag{2}$$

As in par. 3 we divide population in two sub-sets r=RR and $\bar{r} = (NR + UR)$. So it is possible to write:

$$\bar{y} = \frac{1}{M} \sum_{h=1}^{H} \sum_{l=1}^{L} \left(\sum_{j=1}^{{}_{r}M_{hl}} y_{hlj} + \sum_{j=1}^{{}_{\bar{r}}M_{hl}} y_{hlj} \right) = \frac{1}{M} \sum_{h=1}^{H} W_h \sum_{l=1}^{L} W_{hl} \left[t_{hl\,r} \bar{y}_{hl} + \left(1 - t_{hl}\right)_{\bar{r}} \bar{y}_{hl} \right] \tag{3}$$

where ${}_{r}M_{hl}$ is the total number of respondent households of size l in the stratum h,

$_{\bar{r}}M_{hl}$ is the number of non respondents and unreliable households of size l in the stratum h, and

$$W_h = \frac{M_h}{M}, \quad W_{hl} = \frac{M_{hl}}{M_h}, \quad t_{hl} = \frac{{}_{r}M_{hl}}{M_{hl}}.$$

Let us consider $\hat{\bar{y}}$, a direct estimator based only on respondent households:

$$\hat{\bar{y}} = \frac{1}{M}\sum_{h=1}^{H}\sum_{l=1}^{L}\frac{M_{hl}}{{}_{r}m_{hl}}\sum_{j=1}^{{}_{r}m_{hl}} y_{hlj} \tag{4}$$

Using $\hat{\bar{y}}$ as estimator of $\bar{y}$: it is possible to show that

$$E(\hat{\bar{y}}) = \sum_{h=1}^{H}\frac{W_h}{t_h}\sum_{l=1}^{L} t_{hl} W_{hl}\, {}_{r}\bar{y}_{hl} \tag{5}$$

where:

$$t_h = \frac{{}_{r}M_h}{M_h}.$$

The bias of $\hat{\bar{y}}$, $B(\hat{\bar{y}})$, is given by:

$$B(\hat{\bar{y}}) = E(\hat{\bar{y}}) - \bar{y} \tag{6}$$

$$B(\hat{\bar{y}}) = \frac{1}{t}\sum_{h=1}^{H}\frac{W_h}{t_h}\sum_{l=1}^{L} W_{hl}\left(t_{hl} - t_h\right)_{r}\bar{y}_{hl} + \sum_{h=1}^{H} W_h \sum_{l=1}^{L} W_{hl}\left(1 - t_{hl}\right)\left({}_{r}\bar{y}_{hl} - {}_{\bar{r}}\bar{y}_{hl}\right) = V_I + V_D \tag{7}$$

where:

$${}_{r}\bar{y}_{hl} = \frac{1}{{}_{r}M_{hl}}\sum_{j=1}^{{}_{r}M_{hl}} y_{hlj} \quad \text{and} \quad {}_{\bar{r}}\bar{y}_{hl} = \frac{1}{{}_{\bar{r}}M_{hl}}\sum_{j=1}^{{}_{\bar{r}}M_{hl}} y_{hlj} \tag{8}$$

This result confirms the above mentioned results: two components affect nonrespondent bias. The first of these is independent from the values of the considered variable observed on the nonrespondents. The second one is a function of the differences between values assumed by the considered variable in the respondent and in the nonrespondent populations.

We estimate $B(\bar{y})$ substituting in (6) population values with sample estimates as follows:

$$\hat{t}_{hl} = \frac{{}_r m_{hl}}{m_{hl}} \tag{9}$$

$$\hat{t}_h = \sum_{l=1}^{L} \hat{t}_{hl} W_{hl} = \sum_{l=1}^{L} \frac{{}_r m_{hl}}{m_{hl}} \times \frac{M_{hl}}{M_h} \tag{10}$$

$$\hat{t}_h = \sum_{h=1}^{H} \hat{t}_h W_h \tag{11}$$

$${}_r\hat{\bar{y}}_{hl} = \frac{1}{{}_r m_{hl}} \sum_{j=1}^{{}_r m_{hl}} y_{hlj} \tag{12}$$

Our sample doesn't give us the possibility to estimate ${}_{\bar{r}}\bar{y}_{hl}$, so we use UR data to approximate $\bar{r}$ in the following way:

$${}_{\bar{r}}\hat{\bar{y}}_{hl} = \frac{1}{{}_{\bar{r}} m_{hl}} \sum_{j=1}^{{}_{UR} m_{hl}} {}_{UR} y_{hlj} k_{hl} \tag{13}$$

where:

$$k_{hl} = \frac{{}_{\bar{r}} m_{hl}}{{}_{UR} m_{hl}} \tag{14}$$

5 Results

Table 5 shows the results of this study. At this initial stage only some of the most important food and non-food items have been considered.

The underlying assumptions are restrictive, but a lack of other information sources and the quality of results encourages to go ahead in this direction.

Table 5: 1995 HBS-estimates of bias for some items

Items	Monthly Mean Expenditures (MME)	BIAS (B)	BIAS in percentages (B / MME)×100
Bread	43,746	260.75	0.60
Meat	42,017	95.22	0.23
Fruit and vegetables	33,439	359.85	1.08
Electricity	46,640	-149.83	-0.32
Gas	56,552	-402.65	-0.71
Telephone	55,543	139.07	0.25

For all items it is possible to observe small absolute bias values which surpass 1% of monthly mean expenditures only in the fruit and vegetables case.

Concerning food items, the method highlights an overestimation of expenditures evaluated only on the basis of respondents. Expressed as a percentage, this overestimate ranges from 0.23 for meat, to 1.08 for fruit and vegetables.

For non-food items the results reveal an underestimation for gas and electricity (respectively -0.71 and -0.32), and an overestimation for telephone (0.25).

It is our intention to repeat this exercise for 1996 HBS data increasing the number of items considered.

Further studies will include tests of robustness of methods and the possibility of specifying a probabilistic model that, starting from the results of a given number of items, allows us to generalise results so as to estimate bias for the complete set of expenditures.

References

Bethlem, J.G. and Kersten, H.M.P. (1981). The nonresponse problem in the Netherlands. Paper presented at the U.N. Conference of European Statisticians, Meeting relating to problems in household Survey, Geneve

Cochran, W.G. (1977). Sample Techniques. J.Wiley: New York

Hansen, M.H. and Hurwitz, W.N. (1946). The Problem of Nonresponse in Sample Survey. Journal of The American Statistical Association, 41, pp. 517-529

Holt, D. and Smith, T.M.F. (1979). Post-stratification. J.R. Statist. Soc. A., 142, pp. 33-46

ISTAT (1994). Indagine sui consumi delle famiglie - Anno 1993. Collana di informazione, 22, Roma

Kalton, G. and Kasprzyk, D. (1986). The Treatment of Missing Survey Data. Survey Methodology, 12, 1, June 1986, Statistics Canada

Lessler, J.T. and Kalsbeek, W.D. (1992) Non Sampling Error in Surveys. Wiley Series in probability and mathematical statistics, chapter 8

Madow, W.G., Nisselson, H. and Olkin, I. (eds.) (1983). Incomplete Data in Sample Surveys, Vol. 1. Academic Press: New York

Madow, W.G., Olkin, I. and Rubin, D.B. (eds.) (1983). Incomplete Data in Sample Survey, Vol. 2. Academic Press: New York

Mopsik, J.A. and Dippo, C.S. (1984). The adjustment for the consumer Expenditure Interviewed Survey. Proc. Comp. Sect., A.S.A

Platek, R. and Gray, G.B. (1986). On the Definition of Response Rate. Survey Methodology, 12, 1. Statistics Canada

Rubin, D.B. (1983). Conceptual Issues in the Presence of Nonresponse. Incomplete Data in Sample Survey, Vol. 2. Academic Press: New York

Thomsen, I. (1973). A note on the efficiency of weighting subclass means to reduce the effects of non response when analyzing survey data. Statisstisk Tidskrift, 4