

CATI Surveys: Nonparametric Methods to Adjust Estimators for Nonresponse

Indagini CATI: Metodi Non-Parametrici per Correggere le Stime in Presenza di Non Risposta

Andrea Giommi (*), Monica Pratesi (**), Emilia Rocco (*)

(*) Dipartimento di Statistica “G. Parenti” Università di Firenze

(**) Dipartimento di Statistica e Matematica Applicata all’Economia Università di Pisa

Riassunto: Nell’articolo si propone di migliorare la qualità delle stime in presenza di non risposta, nelle indagini telefoniche, mediante l’uso di probabilità individuali di contatto e probabilità individuali di risposta. Tali probabilità vengono stimate mediante un metodo non parametrico e utilizzando informazioni in parte disponibili da lista e in parte provenienti dall’ *automated call scheduling* del CATI. La metodologia proposta viene applicata a dati provenienti da un’indagine reale e confrontata con altri due metodi di stima delle probabilità individuali contatto e di risposta.

Keywords: nonresponse adjustment, telephone survey, automated call scheduling, similarity matrix.

1. Introduction

In this paper we propose to face nonresponse, adjusting the survey estimators by means of individual contact probabilities (ICP) and individual response (collaboration) probabilities (IRP). Obviously, both ICP and IRP are not known and must be estimated from available data. As we refer to CATI surveys in which interviews are managed by an automated call scheduling (ACS), we propose to estimate both ICP and IRP adding the information provided by the ACS to those available from the list. Further we propose a nonparametric approach for estimating these probabilities. This method, that generalizes those advanced by Giommi (1987), is described in section two. In section three, we present an application of the proposed method to real CATI survey data and conclude with a brief discussion of our results.

2. Nonparametric estimation of contact and response probability

Let U be a target population of N units labelled k ($k=1, \dots, N$) and Y a study variable of which we want to estimate the mean. A randomised scheme is adopted for selecting a sample s of fixed size n . Let $c \subseteq s$ be the set of n_c units contacted by the telephone interview and $r \subseteq c$ be the set of n_r contacted units that accept to cooperate at the interview. As we consider an interview the final result of a three step process - selection in the sample, contact and cooperation - we introduce three distribution frames consisting of sample design, contact probability distribution and response probability distribution. The ICP and IRP are derived from the last two distributions as follows:

- (1) ICP: given the sample s and the derived inclusion probabilities π_k , the probability of contacting the set c of selected units is $p(c|s)$. We assume that $p(c|s) \geq 0$ for all $c \in C_s$ and $\sum_{C_s} p(c|s) = 1$, where C_s is the set of all possible sub-sample of contacted units given the selected sample s . Letting $C_{ks} = \{c : k \in c \subseteq s\}$ the ICP, say p_k , is defined as $p_k = \sum_{C_{ks}} p(c|s)$.
- (2) IRP: given the set c of contacted units the probability of achieving the set r of respondents is $q(r|c)$. We assume that $q(r|c) \geq 0$ for all $r \in R_c$ and $\sum_{R_c} q(r|c) = 1$, where R_c is the set of all possible sub-sample of respondent units given the contacted units c . Analogously to the ICP, the IRP is defined as $q_k = \sum_{R_{kc}} q(r|c)$, with $R_{kc} = \{r : k \in r \subseteq c\}$.

We consider only the case where units are contacted and respond independently from each other and from s .

Of course p_k and q_k are unknown and must be estimated from available information.

The idea at the basis of our estimation of p_k and q_k is that couples of units that have similar or identical values for a set of variables, reasonably connected with the response mechanism, may be associated with similar contact and response probabilities. More precisely, our proposal is to estimate the ICP of unit k (for each $k \in s$) as contact rate (i.e. proportion of contacted units) in the group of nearest neighbours of the same unit. An analogous method will be used even for IRP. The nearest neighbours are defined on the basis of the similarity coefficient proposed by Gower (1971). It is defined as:

$$s_{kk'} = \frac{\sum_{l=1}^p w_{kk'l} s_{kk'l}}{\sum_{l=1}^p w_{kk'l}}, \text{ where } s_{kk'l} \text{ is the similarity between the } k\text{th and } k'\text{th}$$

individual as measured by the l th variable and $w_{kk'l}$ is 1 or 0 depending on whether or not the comparison is considered valid for the l th variable. We chose to assign value 1 to every $w_{kk'l}$ and use for all the variables $s_{kk'l} = 1 - |X_{kl} - X_{k'l}| / R_l$, where X_{kl} and $X_{k'l}$ are the two individuals' values for variable l , and R_l is the range of the same variable.

Estimated ICP and IRP are then used for adjusting the Horvitz and Thompson type estimator of the mean that becomes:

$$\hat{\bar{Y}} = \frac{\sum_{k \in r} y_k / \pi_k \hat{p}_k \hat{q}_k}{\sum_{k \in r} 1 / \pi_k \hat{p}_k \hat{q}_k}$$

3. Application of proposed methods to a survey on student satisfaction for library services in the University of Florence

The methods described in the previous section have been applied to a real CATI survey carried out by the Department of Statistics of the University of Florence in order to evaluate the satisfaction of the students for the services of the university libraries. The objective population consists of all the students who borrowed at least one book during the year 2000. The size, N , of the population is equal to 27746. A sample of 472 students was selected using a random stratified design with strata representing by the

different libraries. Only 277 students were contacted and 259 completely respond to the questionnaire.

We have defined the contact by an indicator variable which assumes value 1 when the interviewer talked exactly with the sampled student and 0 in all the other cases. In order to estimate the ICP, we observed that only five variables showed a significant association with the variable contact. Three were derived from the ACS: i) the number of appointments made for the interview (with values 0, 1, 2, 3); ii) the number of calls with result: busy, free, answering machine (with values 0, 1, 2, 3); iii) a dichotomous variable with value 1 if the student to be interviewed had changed his telephone number (some of these were still found) and 0 otherwise. Two were list variables: i) age (at last birthday); ii) residence region (with value 1 if Tuscany and 0 otherwise).

For the estimation of the IRP, given the small number of nonresponses, we found only two variables, showing significant association with response: one from the ACS, the total number of calls (coded with value 1 if this number was less than three, and 0 otherwise) and one list variable, the total number of borrowed books (coded with value 1 if the number was less than seven and 0 otherwise).

The set of neighbours needed to calculate the probability of contact and response for each contacted and respondent student has been defined as follows. We calculate $s_{kk'l}$ and we choose the first 60 units with the same $s_{kk'l}$ around the contacted unit (or response unit). If in addition to the 60th neighbour there are others having the same value $s_{kk'l}$ we included even those in the estimation set. The resulting estimation sets range from 60 to 118 units for the ICP and from 60 to 217 for the IRP.

Estimated ICP and IRP have been used for adjusting the Horvitz and Thompson estimator of the mean. To evaluate its performance we estimated the average number of books borrowed by the students during the year 2000 ($\bar{Y}_1$) and the mean of the student ages ($\bar{Y}_2$) which are known for the whole population. Furthermore, for each of these variables, we compared this estimate with two other estimates: the one adjusted with ICP and IRP both estimated by logistic model, and the one adjusted with ICP and IRP both estimated as overall contact and response rates (the estimator based only on the respondents). The results for these control variables are reported in table 1 and 2, where: *NP* denotes the proposed method, with non parametric estimation of the ICP and IRP used for adjusting the estimates; *Log* denotes the method where the ICP and IRP are estimated by logistic model; *R* is the method based only on the respondents, where the proportions of contact and response are assumed to be constant among the students and equal to the contact and response rates.

For each method we estimated the mean, the ratio of the estimated mean to the real value (which are, respectively, 6.28 for the books and 24.97 for the age), the standard error of the estimates $se(\cdot)$ evaluated empirically by the following methods: dependent random group method, splitting the sample into six sub-samples (with the same design of the full sample) and using the full sample estimate as reference values: we called $se_{rg}(\cdot)$ the corresponding index; jackknife method based on sub-samples obtained omitting in turn one unit of the full sample and using the full sample estimate for computing $se_{jk}(\cdot)$; the bootstrap method, based on 1000 bootstrap samples drawn from an artificial population built on the basis of the full sample data and assumed to mimic the real, though unknown, population under study. We used the mean of the 1000 bootstrap samples $se_{bs}(\cdot)$.

Table 1: *Estimates of average number of books borrowed by the students, under three adjustment methods for noncontact and nonresponse*

<i>Meth ods</i>	$\hat{\bar{Y}}_1$	$\hat{\bar{Y}}_1/\bar{Y}_1$	$se_{rg}(\hat{\bar{Y}}_1)$	$se_{jk}(\hat{\bar{Y}}_1)$	$se_{bs}(\hat{\bar{Y}}_1)$
<i>NP</i>	6.13	0.976	0.494	0.460	0.434
<i>Log</i>	5.97	0.951	0.524	0.497	0.436
<i>R</i>	6.11	0.973	0.447	0.431	0.412

Table 2: *Estimates of average student age under three adjustment methods for noncontact and nonresponse*

<i>Meth ods</i>	$\hat{\bar{Y}}_2$	$\hat{\bar{Y}}_2/\bar{Y}_2$	$se_{rg}(\hat{\bar{Y}}_2)$	$se_{jk}(\hat{\bar{Y}}_2)$	$se_{bs}(\hat{\bar{Y}}_2)$
<i>NP</i>	24.90	0.997	0.433	0.303	0.301
<i>Log</i>	25.01	1.002	0.476	0.375	0.416
<i>R</i>	24.59	0.985	0.395	0.271	0.264

The nonparametric approach we have proposed in the present work for the estimation of both contact and response probabilities is very simple and based on logical foundations. Simplicity lies in the fact that once defined an opportune measure of similarity between units, probabilities of contact and response are estimated simply by ratios inside classes of units. Logicity of the method relates on the fact that the sample unit for which the estimation is carried out is always more similar to the units belonging to its weighting class than to all the other units outside the class. The results deriving from the application are favourable to our method with respect to the *Log* one, even if tables 1 and 2 show that there are no substantial differences among the estimates adjusted with the different techniques. The estimates based on respondents only show less variability than the other two, but they also present a higher difference with the true value (this confirms their usual bias). Furthermore, as said above, the corresponding assumption of constant probabilities of contact and response or, equivalently, completely random noncontact and nonresponse is often unrealistic. The closeness of the estimates obtained on the respondents only and those obtained adjusting the estimates by both the other two methods may suggest that, in our case study, the probabilities of contact and response are weakly related to the individual features of the students. The weakness of this relation is also confirmed by the low variability of the response probabilities (the standard deviation of the IRP is 0.016 when they are estimated by nonparametric method and 0.045 when they are estimated by logistic one).

References

- Giommi A. (1987), "Nonparametric methods for estimating individual response probabilities", *Survey Methodology*, si3, 2, pp. si27-si34.
- Gower (1971), "A general coefficient of similarity and some of its properties", *Biometrics*, 27, pp. 857-872.
- Groves R. M., Lyberg L. E. (1988), "An overview of nonresponse issues in telephone surveys", In Groves R. M. et Al. (eds) *Telephone Survey Methodology*, J. Wiley and Sons, New York.