

Estimating Measurement Error through a Reinterview Survey Using a Different Data Collection Technique

Stima dell'Errore di Misura Mediante una Indagine di Controllo basata su una differente Tecnica di Raccolta dei Dati

Piero Demetrio Falorsi

Alessandro Pallara

ISTAT

Via Depretis 74/b, 00184

Rome (Italy)

e-mail: falorsi@istat.it

e-mail: pallara@istat.it

Aldo Russo

Department of Political Institutions and
Social Sciences

University Roma Tre; Rome

Riassunto: Questo lavoro propone un metodo dei momenti per la stima della componente semplice e di quella correlata dell'errore di misura. Si affronta il caso in cui per la stima degli errori di misura si possa disporre dei dati di un'indagine di controllo in cui la seconda misurazione sia stata effettuata con una tecnica di rilevazione differente da quella adottata nella prima indagine e che quindi le due misurazioni (quella della prima indagine e quella dell'indagine di controllo) siano soggette a una differente variabilità dell'errore di misura. Tale contesto sembra più realistico di quello proposto correntemente nella letteratura sull'argomento, in quanto spesso nelle condizioni concrete i vincoli di costo ed operativi rendono impossibile il fatto che la seconda misurazione sia effettuata esattamente con la medesima tecnica adottata nella prima indagine.

Keywords: Measurement errors; Simple and Correlated Measurement Variance.

1. Introduction

One important factor affecting the accuracy of the survey estimates is represented by the component of measurement errors which can be originated during the collection stage by the respondents, the interviewers and the relationships among them. As a tool to estimate the impact of measurement errors, a *reinterview* on a subsample of respondents can be used in order to gather repeated measures and apply appropriate models (Cicchitelli *et al.*, 1992; cap.XIII). When using the approach based on *reinterviews* and two measures on the same respondents are gathered, the following hypotheses are usually considered (Forsman, 1989; Hansen *et al.*, 1961): (i) the two measures are modelled as random variables; (ii) the repeated measures on the same unit are independent; (iii) the measurement conditions for the two occasions are identical or as close to identical as possible (Särndal *et al.*, 1992; pp. 614-616); this imply that the random variables which are associated to the two measurements are subjected to the *same measurement error variance*.

In this paper, keeping the assumption of the independence of the two measurements, we attempt to estimate the *simple and correlated variance components* when a reinterview

is available on a subsample of respondents, but the two measures can not be considered as being collected under the same conditions, and are really subject to *different measurement error variance*. This seems a more realistic condition when, in the actual cases, it is impossible to assure the same measurement conditions in the two interviews; indeed, operational and budget considerations often suggest to adopt a different survey mode in the second interview. For example, in the *Italian Census on Agriculture* the interview has been conducted by a *face to face* technique; while for budget constraints, reinterview program for the assessment of the quality of census data has been based on a validation survey which has been conducted through a Computer Assisted Telephone Interview (CATI) technique.

2. The statistical model

Consider the following conditions: (i) a finite population P is constituted by N units; (ii) an original random sample s (survey I_1) of size n is drawn from the population P with a sampling design $p(\cdot)$, with p_k representing the probability that element k will be included in the sample and p_{kl} representing the probability that both of the elements k and l will be included in the sample; (iii) for each element $k \in s$, $y_{k,1}$ denotes the observed value of the variable y of interest; a second phase sample r ($r \subseteq s$) (survey I_2) of size n_r is drawn according to the design $p(\cdot|s)$, such that $p(r|s)$ is the conditional probability of choosing r . The inclusion probabilities under this design are denoted $p_{k|s}$ and $p_{kl|s}$ for elements k and $l \in s$; (iv) for each element $k \in r$, $y_{k,2}$ denotes the observed value of the variable y of interest. The measurement model m is specified as follows:

$$y_{k,t} = m_k + e_{k,t} \quad (t=1,2) \quad (1)$$

where m_k is the *true value* and $e_{k,t}$ is the *erratic* component.

Given the general survey conditions of I_1 and of I_2 , the expectation values $E_m(\cdot)$ under the model (1) are:

$$\begin{aligned} E_m(y_{k,t}) &= m_k && \text{for } k \in s \text{ if } t=1 \text{ and } k \in r \subseteq s \text{ if } t=2 \\ E_m(y_{k,t} - m_k)^2 &= s_{k,t}^2 && \text{for } k \in s \text{ if } t=1 \text{ and } k \in r \subseteq s \text{ if } t=2 \\ E_m[(y_{k,t} - m_k)(y_{l,t} - m_l)] &= s_{kl,t} && \text{for } k, l \in s \text{ if } t=1 \text{ and } k, l \in r \subseteq s \text{ if } t=2 \\ E_m[(y_{k,1} - m_k)(y_{l,2} - m_l)] &= 0 && \text{for } k, l \in r \subseteq s. \end{aligned}$$

It is important to note that in the above model both the values m_k and $s_{k,t}^2$, are related to the specific unit $k \in P$, while $s_{kl,t}$ is a value related to the specific couple $(k, l) \in P$, independently by the selected sample in the I_t ($t=1$ or 2) survey.

The objective of the survey I_1 is to estimate the population total of the *true values*:

$t_y = \sum_P m_k$, where we denote with $\sum_P$ the sum over the N units of the population P . An estimate of t_y may be obtained by the Horvitz-Thompson estimator

$$\tilde{t}_y = \sum_s \frac{y_{k,1}}{p_k}. \quad (2)$$

3. Decomposition of the variance

As a measure of the accuracy of the estimator (2), we use the variance of $\tilde{t}_y$ defined with respect to *design* $p(\cdot)$ and *model* m jointly

$$V_{pm}(\tilde{t}_y) = E_{pm}[(\tilde{t}_y - t_y)^2] \quad (3)$$

where the indexes p and m define the operators E and V with respect to the sample design and to the model.

The variance term $V_{pm}(\tilde{t}_y)$ can be decomposed as follows

$$V_{pm}(\tilde{t}_y) = E_p[V_m(\tilde{t}_y|s)] + V_p[E_m(\tilde{t}_y|s)].$$

The *measurement variance*, given by $E_p[V_m(\tilde{t}_y|s)] = V_{A1} + V_{A2}$, may be expressed by the sum of two terms, where

$$V_{A1} = \sum_p s_{k,1}^2 + \sum_{k \neq l} \sum_p s_{kl,1} ; V_{A2} = \sum_p \frac{(1-p_k)}{p_k} s_{k,1}^2 + \sum_{k \neq l} \sum_p \frac{(p_{kl} - p_k p_l)}{p_k p_l} s_{kl,1}.$$

The component $V_p[E_m(\tilde{t}_y|s)] = V_B$, denoting the *sampling variance*, is expressed by

$$V_p[E_m(\tilde{t}_y|s)] = V_B = \sum_k \sum_l \sum_p \frac{p_{kl} - p_k p_l}{p_k p_l} m_k m_l.$$

4. Estimate of the variance

As shown in Särndal *et al.* (1992), in presence of measurement errors, the standard estimator of variance

$$\tilde{V}(\tilde{t}_y) = \sum_k \sum_l \sum_s \frac{p_{kl} - p_k p_l}{p_{kl}} \frac{y_{k,1}}{p_k} \frac{y_{l,1}}{p_l}$$

is biased, being valid the following result $E_{pm}(\tilde{V}(\tilde{t}_y)) = V_{pm}(\tilde{t}_y) - V_{A1}$.

Therefore, for obtaining an unbiased estimate of $V_{pm}(\tilde{t}_y)$, it is sufficient to yield an unbiased estimate of the V_{A1} component. In order to solve this problem, under the statistical model introduced in section 2, it can be shown that a consistent estimate of the vector of five unknown terms $\mathbf{X} = \{X_i\}$ ($i=1, \dots, 5$), where

$$\begin{aligned} X_1 &= \frac{1}{N} \sum_p \left(m_k - \frac{1}{N} \sum_p m_k \right)^2, & X_2 &= \sum_p s_{k,1}^2, & X_3 &= \sum_p s_{k,2}^2 \\ X_4 &= \sum_k \sum_{l \neq k} \sum_p s_{kl,1}, & X_5 &= \sum_k \sum_{l \neq k} \sum_p s_{kl,2}, \end{aligned}$$

may be found as

$$\hat{\mathbf{X}} = \mathbf{A}^{-1} \tilde{\mathbf{b}} \quad (4)$$

in which $\mathbf{A}$ is a matrix of coefficients

$$\mathbf{A} = \begin{bmatrix} 1 & (N-1)/N^2 & 0 & -1/N^2 & 0 \\ 1 & 0 & (N-1)/N^2 & 0 & -1/N^2 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix},$$

and $\tilde{\mathbf{b}} = \{b_i\}$ ($i=1, \dots, 5$), is a vector of *known terms*, with

$$\begin{aligned} \tilde{b}_1 &= \frac{1}{N^2} \left[\sum_s \frac{y_{k,1}^2}{p_k} (N-1) - \sum_k \sum_{l \neq k} \frac{y_{k,1} y_{l,1}}{p_{kl}} \right] \\ \tilde{b}_2 &= \frac{1}{N^2} \left[\sum_r \frac{y_{k,2}^2}{p_{k,2}} (N-1) - \sum_k \sum_{l \neq k} \frac{y_{k,2} y_{l,2}}{p_{kl,2}} \right] \\ \tilde{b}_3 &= \sum_r \frac{1}{p_{k,2}} (y_{k,1} - y_{k,2})^2 \\ \tilde{b}_4 &= \sum_s \frac{y_{k,1}^2}{p_k} + \sum_k \sum_{l \neq k} \frac{y_{k,1} y_{l,1}}{p_{kl}} + \sum_r \frac{y_{k,2}^2}{p_{k,2}} + \sum_k \sum_{l \neq k} \frac{y_{k,2} y_{l,2}}{p_{kl,2}} + \\ &\quad - 2 \sum_k \sum_l \frac{y_{k,1} y_{l,2}}{p_{kl,2}} \\ \tilde{b}_5 &= \sum_r \frac{y_{k,1} y_{k,2}}{p_{k,2}} \left(1 - \frac{1}{N} \right) - \frac{1}{N} \sum_k \sum_{l \neq k} \frac{y_{k,1} y_{l,2}}{p_{kl,2}} \end{aligned}$$

where $p_{k,2} = p_k p_{k|s}$ and $p_{kl,2} = p_{kl} p_{kl|s}$.

Using expression (4) on the basis of the estimates $\hat{X}_i$ ($i=1, \dots, 5$), it is then possible to obtain an estimate of the variance V_{A1} as

$$\hat{V}_{A1} = \hat{X}_2 + \hat{X}_4. \quad (5)$$

Finally, an estimate of $V_{pm}(\tilde{t}_y)$ may be given by

$$\tilde{V}_{pm}(\tilde{t}_y) = \tilde{V}(\tilde{t}_y) + \hat{V}_{A1} \quad (6)$$

References

- Cicchitelli G., Herzel A., Montanari G.E. (1992) *Il campionamento statistico*, il Mulino, Bologna.
- Forsman G. (1989) Early Survey Models and Their Use in Survey Quality Work, *Journal of Official Statistics*, Vol. 5 No. 1, 41-55.
- Hansen M., Hurwitz W., Bershad M. A. (1961) Measurements Errors in Census and Surveys, *Bullettin of the International Statistical Institute*, 38:2, 359-374.
- Särndal C. E., Swensson B., Wretman J. (1992) *Model Assisted Survey Sampling*, Springer-Verlag, New York.