

Jackknife Linearization Variance Estimator for Poststratified Estimators in the Presence of Unit Nonresponse

Stimatore jackknife linearizzato della varianza per stimatori post-stratificati in presenza di unità non rispondenti

Paolo Righi, Fabrizio Solari

ISTAT MPS/A, Via A. Depretis, 74/B, 00184, Roma,

parighi@istat.it, solari@istat.it

Riassunto: Il problema delle unità non rispondenti in una indagine campionaria è generalmente risolto con una procedura di correzione dei pesi campionari che prevede la partizione del campione selezionato in sottoclassi o “celle di aggiustamento” al cui interno la probabilità di risposta delle unità campionarie è supposta costante. Nel caso in cui le celle di aggiustamento non coincidano con eventuali post-strati, gli stimatori tradizionali della varianza, basati sul metodo della linearizzazione di Taylor, spesso conducono a sottostime. Al contrario, lo stimatore jackknife linearizzato della varianza risulta consistente sotto deboli condizioni. In questo lavoro viene proposta una generalizzazione di tale stimatore che garantisce maggiore stabilità alla stima della varianza nel caso in cui si presentino poche unità per cella.

Keywords: Unit nonresponse, Two-phase sampling, Poststratified estimator, Jackknife linearization variance estimator.

1. Introduction

Large-scale surveys are usually affected by nonresponse. A general classification defines two types of nonresponse: unit nonresponse and item nonresponse. The former appears when all responses about a sampled unit are not available, while the latter arising when some but not all responses are available. In presence of unit nonresponse, considered in this paper, generally, an unbiased estimator becomes biased and its variance increases. One way to eliminate or to reduce bias is to compensate for the unit nonresponse by a weighting adjustment procedure. This procedure is based on the assumption that the response behaviour is stochastic (Cochran, 1977) and that the population is composed of a set of disjoint subpopulations in each of them all units have the same response probability and respond independently of each other. Models based on these hypotheses are known as “uniform response mechanism within subpopulation model” (Oh and Scheuren, 1983) and “response homogeneity group model” (Särndal and Swensson, 1987). In practice, the procedure divides the sample into adjustment cells formed by the cross-classification from a set of auxiliary variables known for all the sampled units, respondents and nonrespondents, thought to be correlated with sample response (Little, 1986). Assuming that uniform response mechanism in each cell holds, separate weighting adjustments are applied within each adjustment cell so that

the respondents represent the nonrespondents. The nonresponse adjustment factor is equal to the inverse of the estimated response probability in each cell. Alternative methods for performing the nonresponse adjustment are described in the literature (see for example Rizzo *et al.*, 1996; Särndal *et al.*, 1992).

2. The poststratified estimator in presence of unit nonresponse

Suppose that we have a stratified multistage design with H strata and N_h primary sampling units (PSUs) in stratum h . The i -th PSU in stratum h , denoted with (hi) , contains M_{hi} secondary sampling units (SSUs). A first stage sample of n_h PSUs is drawn from each stratum and m_{hi} SSUs are sampled if (hi) is selected. We denote the k -th sampled SSU in the i -th PSU of the h -th stratum with (hik) . Assume, also, that each SSU belongs to a poststratum and that the total number of SSUs in the c -th poststratum, ${}_cM$, is known. Finally, we suppose that the sample suffers from unit nonresponse. Applying the weighted adjustment procedure described above, Yung and Rao (2000) used the following nonresponse adjustment factor for the p -th cell

$$d_p = \frac{\sum_{(hik)} w_{hik} h_{hik}^p}{\sum_{(hik)} w_{hik} a_{hik} h_{hik}^p},$$

where w_{hik} is the design weight, a_{hik} is the response indicator and h_{hik}^p is the adjustment cell indicator.

Then, the poststratified estimator of the total Y in the presence of unit nonresponse is

$$\hat{Y}_{PS} = \sum_c {}_cM \frac{\sum_p d_p {}_c\hat{Y}_p^r}{\sum_p d_p {}_c\hat{M}_p^r}, \quad (1)$$

where ${}_c\hat{Y}_p^r = \sum_{(hik)} w_{hik} a_{hik} {}_c\mathbf{d}_{hik} \mathbf{h}_h^p y_{hik}$, ${}_c\hat{M}_p^r = \sum_{(hik)} w_{hik} a_{hik} {}_c\mathbf{d}_{hik} \mathbf{h}_h^p$, ${}_c\mathbf{d}_{hik}$ is the poststratum indicator and y_{hik} is the observed value of the variable of interest.

The nonresponse adjustment for the p -th cell in (1) uses only the information from units belonging to the same adjustment cell p . However, the sample size in some of the cells could be equal to zero or not large enough to ensure the stability of the nonresponse adjustment factor d_p (for example Rizzo *et al.*, 1996, suggest less than 25 units) and, therefore, of the estimator $\hat{Y}_{PS}$ (Oh and Scheuren, 1983).

In order to solve this problem we introduce two-phase sampling estimator approach. Without loss of generality, we suppose that each adjustment cell results from combinations of only two auxiliary variables A_1 and A_2 , assuming, respectively, S and T levels, so that the total number of adjustment cells is $P = S \times T$. Having the same sampling information utilized by Yung and Rao, we obtain the following poststratified estimator in the presence of unit nonresponse

$$\hat{Y}_{PSTP} = \sum_c {}^R \hat{Y}_r / {}^R \hat{M}_r = \sum_c M \frac{\sum_p g_p f_p^{-1} {}^R \hat{Y}_p^r}{\sum_p g_p f_p^{-1} {}^R \hat{M}_p^r}, \quad (2)$$

where ${}^R \hat{Y}_r$ and ${}^R \hat{M}_r$ are, respectively, the two-phase sampling generalized regression estimator of ${}_c Y$ and ${}_c M$; f_p is the response rate given by the ratio of the number of respondents, r_p , and the total number of SSUs, m_p , in the p -th cell; g_p the two-phase regression correction factor given by the following expression

$$\begin{aligned} g_p &= 1 + \left[\sum_p \hat{X}_p - \sum_p f_p^{-1} \hat{X}_p^r \right] \left(\sum_p f_p^{-1} \hat{A}_p^r \right)^{-1} \mathbf{x}_p, \\ &= 1 + \left(\hat{X} - \hat{X}_r \right)' \hat{A}_r^{-1} \mathbf{x}_p, \end{aligned} \quad (3)$$

with $\hat{X}_p = \sum_{(hik)} w_{hik} h_{hik}^p \mathbf{x}_{hik}$, $\hat{X}_p^r = \sum_{(hik)} w_{hik} a_{hik} h_{hik}^p \mathbf{x}_{hik}$, $\hat{A}_p^r = \sum_{(hik)} w_{hik} a_{hik} h_{hik}^p \mathbf{x}_{hik} \mathbf{x}_{hik}'$ and being the vector $\mathbf{x}_p$, related to the p -th cell, as the explanatory variables vector in an ANOVA model with two factors, A_1 and A_2 . Simple algebra shows that the estimator (1) can be viewed as a particular case of (2) when $\mathbf{x}_p$ is defined in the same way as the explanatory variables vector of a saturated ANOVA model. However, in (3) other choices of $\mathbf{x}_p$ are possible. For instance, in order to solve the problem of little information in the adjustment cells, we can use the vector $\mathbf{x}_p$ related to the two-way ANOVA model without interaction parameters. In fact, in this case the nonresponse adjustment factor for the p -th cell, obtained by the combination of level s of A_1 and level t of A_2 , utilizes all the SSUs belonging to level s or t . This approach leads to good variance properties and, if the interaction between A_1 and A_2 is moderate, small bias, producing an estimator with mean square error less than the estimator (1).

3. The Jackknife Linearization Variance Estimator

Weighting adjustment procedures should be taken into account defining the variance estimator. In (1) or (2) when the adjustment cells cut across the poststrata, applying the standard variance estimator, as Taylor linearization variance method, leads to serious underestimation if the nonresponse rate is appreciable. Yung and Rao (2000) proposed a variance estimator for the poststratified estimator in (1), and in general for the generalized regression estimator, in order to handle this general nonresponse framework. The variance estimator is obtained by linearizing the jackknife variance estimator. Yung and Rao (1996) showed that the linearized form performed as well as the full jackknife and it is consistent under mild conditions. Moreover, this approach is less demanding computationally than full jackknife not requiring recalculation of the weights for each deletion. Now, we compute the jackknife linearization variance estimator $v_{JL}(\hat{Y}_{PSTP})$ of (2)

$$v_{JL}(\hat{Y}_{PSTP}) = \sum_h \frac{n_h}{n_h - 1} \sum_i (e_{hi}^r - \bar{e}_h^r)^2,$$

where

$$\begin{aligned} e_{hi}^r = & \sum_c \sum_p \sum_k f_p^{-1} g_{pc} b w_{hik} a_{hik} d_{hik} h_{hik}^p (y_{hik} - {}^R\hat{Y}_r / {}^R\hat{M}_r) \\ & + \sum_c \sum_p \sum_k c b w_{hik} \left[(1 - a_{hik} \sum_p f_p^{-1} h_{hik}^p) - (\hat{\mathbf{X}} - \hat{\mathbf{X}}_r)' \hat{\mathbf{A}}_r^{-1} a_{hik} \mathbf{x}_{hik} \sum_p f_p^{-1} h_{hik}^p \right] \\ & \times (\mathbf{x}_{hik}' \hat{\mathbf{A}}_r^{-1} \mathbf{x}_p) \left[f_p^{-1} \hat{Y}_p^r - ({}^R\hat{Y}_r / {}^R\hat{M}_r) f_p^{-1} \hat{M}_p^r \right] \\ & + \sum_c \sum_p \sum_k c b r_p^{-1} \left\{ 1 - \left[f_p^{-1} \hat{\mathbf{X}}_p^r + (\hat{\mathbf{X}} - \hat{\mathbf{X}}_r)' (\mathbf{I} - f_p^{-1} \hat{\mathbf{A}}_r^{-1} \hat{\mathbf{A}}_p^r) \right] \hat{\mathbf{A}}_r^{-1} \mathbf{x}_p \right\} \\ & \times \left[h_{hik}^p (1 - a_{hik} f_p^{-1}) \right] \left[f_p^{-1} \hat{Y}_p^r - ({}^R\hat{Y}_r / {}^R\hat{M}_r) f_p^{-1} \hat{M}_p^r \right] \end{aligned} \quad (4)$$

$\bar{e}_h^r = (1/n_h) \sum_i e_{hi}^r$, ${}_c b = {}_c M / {}^R\hat{M}_r$ is the poststratification adjustment factor and $\mathbf{I}$ is the identity matrix. Although the expression of the residual appears complicated to compute, it only requires the calculation of few quantities, turning out to be less time spending than the full jackknife procedure. Finally, e_{hi}^r reduces to the first term of the formula (4) in the full response case or when the weighting adjustment cells are the poststrata. Moreover, we obtain again the residual term computed by Yung and Rao as a particular case of (4).

References

- Cochran W.G. (1977), *Sampling Techniques*, Wiley, New York.
- Little R.J.A. (1986) Survey Nonresponse Adjustments for Estimates of Means, *International Statistical Review*, 54, 139-157.
- Oh H.L., Scheuren F. (1983) Weighting adjustments for unit nonresponse, in: *Incomplete Data in Sample Surveys*, Madow W.G., Olkin I., Rubin D., (Eds.), Academic Press, New York, 2, 143-184.
- Rizzo L., Kalton G., Bick J.M. (1996) A Comparison of Some Weighting Adjustment Methods for panel Nonresponse, *Survey Methodology*, 22, 43-53.
- Särndal C.E., Swensson B. (1987) A General View of Estimation for Two Phases of Selection with Application to Two-Phase Sampling and Nonresponse, *International Statistical Review*, 55, 279-294.
- Särndal C.E., Swensson B., Wretman J. (1992) *Model Assisted Survey Sampling*, Springer-Verlag, New York.
- Yung W., Rao J.N.K. (1996) Jackknife Linearization Variance Estimators Under Stratified Multi-Stage Sampling, *Survey Methodology*, 22, 23-31.
- Yung W., Rao J.N.K. (2000), Jackknife Variance Estimation Under Imputation for Estimators Using Poststratification Information, *Journal of the American Statistical Association*, 95, 903-915.