Nonresponse Bias Reduction Integrating Calibration Constraints with Variance and Covariance Structure of Auxiliary Variables

Attenuazione della Distorsione Originata dai Non Rispondenti Integrando i Vincoli di Calibrazione con la Struttura di Varianza e Covarianza delle Variabili Ausiliarie

Antonio Pavone
ISTAT, pavone@istat.it

Riassunto: Per ridurre gli effetti distorsivi nelle stime finali causati dalle mancate risposte totali, Lundström e Särndal (1999) suggeriscono d'impiegare lo stimatore di calibrazione come strumento prevalente nella fase di stima. Tuttavia, se il campione dei rispondenti differisce sistematicamente dai non rispondenti, la procedura di modifica dei pesi campionari, indicata da questi autori, non corregge a sufficienza l'effetto distorsivo sulle stime, dovuto all'autoselezione dei rispondenti. In questo articolo si propone una modifica dello stimatore di calibrazione che incorpori, tra i vincoli da soddisfare, anche la struttura di varianza e covarianza conosciuta delle variabili ausiliarie. Tale variante sembra attenuare più efficacemente la distorsione delle stime e può essere considerata una valida alternativa, quando risulta arduo modellizzare la mancata risposta.

Keywords: Calibration estimator, selection bias, auxiliary information, nonresponse adjustment.

1. Brief description of the calibration procedure

Consider a fixed finite population of size N called *target population* $U = (1,...,k,...,N)$. The main aim of a survey is to estimate the total $Y = \sum_{k \in U} y_k$, where y_k, for k units, is the study variable. Supposing sample s is extracted from size n of the *target population*, using a sample design, and the set of respondent units in the sample is indicated as s_r and $(s_r \subseteq s)$. An expression of total estimate Y is $\hat{Y} = \sum_{k \in s_r} y_k d_k \nu_{1k} \nu_{2k}$, where the final weight $w_k = d_k \nu_{1k} \nu_{2k}$ is defined as the product of three elements. The first is the Horvitz-Thompson weight $d_k = 1/\pi_k$, where $\pi_k = \Pr(k \in s)$ for $(k \in U)$ indicates the probability of each unit k to be included in the sample. The second is nonresponse-adjusted Horvitz-Thompson weight $\nu_{1k} = 1/\hat{\theta}_k$ and it is used to correct the respondents' sample by bias selection, where $\hat{\theta}_k = \Pr(k \in s_r | s)$ is the estimate of probability response θ_k for *kth* unit. The third is a post-stratification adjustment factor ν_{2k} able to incorporate new information from the *target population*, not available at the sampling design stage.

Lundström and Särndal 1999 depart from the standard method to estimate of probability response. Using a calibration estimator (Särndal, Swensson, and Wretman 1992; Deville and Särndal, 1992), they derive implicitly the unknown response probabilities. In such way, in one single step, the optimal solution is obtained applying d_k as an initial solution by which new weight w_k is derived. This requires the attribute to be as close as possible to the initial weight and is able to reproduce the known quantities of a set of auxiliary variables. If auxiliary information is available for both sample and population units, then the parameters w_k are found by solving the following optimization problem

$$\begin{cases} \underset{\{w_k\}}{Min}\left[\sum_{k\in s_r}(w_k - d_k)^2 / d_k c_k\right] \\ \sum_{k\in s_r} w_k = N \qquad w_k \in \Re_{++} \\ \sum_{k\in s_r} w_k\, x_{k,p} = \sum_{k\in s} d_k x_{k,p} \qquad p = 1,....,P \\ \sum_{k\in s_r} w_k z_{k,q} = \sum_{k\in U} z_{k,q} \qquad q = 1,....,Q \end{cases}$$

where the minimized distance function between (w_k, d_k) follows an Euclidean metric (for other distance functions see Deville and Särndal, 1992 and Dupont 1994); c_k is a positive factor specified in advance; $\mathbf{z}_k$ is a vector of the total auxiliary variables correlated with study variables and known for the target population; $\mathbf{x}_k$ is the vector of the total auxiliary variables correlated with the response probability and known for all the sample units. Unlike the Lundström and Särndal assessment, the empiric experience shows that, if the respondent sample diverges systematically from the non respondents, a conventional calibration approach cannot reduce the nonresponse bias as much as possible, if not including an explicit modellization of the response probability.
To improve the performance of the calibration estimator, in this article is suggested to take into account the variance-covariance structure of the known auxiliary variables totals, in the equality constraints

$$\begin{cases} \sum_{k\in s_r} w_k\, x_{k,p}\, x_{k,p'} = \sum_{k\in s} d_k x_{k,p}\, x_{k,p'} \qquad p = 1,...,P;\, p' = 1,...,P \\ \sum_{k\in s_r} w_k\, z_{k,q}\, z_{k,q'} = \sum_{k\in U} z_{k,q}\, z_{k,q'} \qquad q = 1,...,Q;\, q' = 1,...,Q \end{cases}$$

Such options can be considered a valid alternative, when it is difficult to model the functional form of relations between the nonresponse unit and a set of auxiliary variables. A limited simulation study, described in section 2, shows that, adding the joint and quadratic effects to the calibration requirements, the nonresponse bias of Y estimate is removed more effectively than using only the main effect of auxiliary variables.

2. An examination of nonresponse unit bias

For simulation purposes it is assumed that the estimated variable is *Y*, the auxiliary variables observed $(\mathbf{x}_k = \mathbf{z}_k)$ are *X1* and *X2*, and *Y*, *X1* and *X2* are closely associated. The artificial population is composed of 10000 units subdivided into subsets of 5000 units, indicated as DUM0 and DUM1. A restriction is imposed: each subset is heterogeneous in terms of mean, variance and covariance. The table below displays some key statistics for each subset separately.

Table 1: *Statistical key values for the population characteristics*

Descriptives

DUM0			DUM1		
Variables	**Mean**	**St. Deviation**	**Variables**	**Mean**	**St. Deviation**
Y	236.557	166.611	*Y*	562.602	246.157
X1	255.358	166.898	*X1*	451.382	208.801
X2	497.311	288.075	*X2*	499.846	258.064

Test of Equality of Error Variances and Covariance Matrices

Levene's Test					Box's Test				
Variables	**F**	**df1**	**df2**	**Sig.**	**Box's M**	**F**	**df1**	**df2**	**Sig.**
Y	1239.270	1	9998	0.000	10782.990	1796.581	6	724238520	0.000
X1	446.288	1	9998	0.000					
X2	63.761	1	9998	0.000					

Correlation Matrix for DUM0 and DUM1

DUM0				DUM1			
	Y	*X1*	*X2*		*Y*	*X1*	*X2*
Y	1.000	0.982	0.973	*Y*	1.000	0.992	0.988
X1	0.982	1.000	0.982	*X1*	0.992	1.000	0.991
X2	0.973	0.982	1.000	*X2*	0.988	0.991	1.000

Least square linear regression for DUM0 and DUM1. Dependent variable = *Y*

DUM0 (Adjusted R^2 =0.967)				**DUM1** (Adjusted R^2 =0.986)			
Predictor Variables	**Coefficients**	**Std. error**	**Student's t**	**Predictor Variables**	**Coefficients**	**Std. error**	**Student's t**
constant	-22.046	0.932	-23.646	*constant*	49.350	1.151	42.859
X1	0.767	0.014	56.608	*X1*	0.827	0.014	57.819
X2	0.126	0.008	16.096	*X2*	0.280	0.012	24.156

To draw 250 SRS samples of dimension n = 1000, it is assumed that the nonresponse mechanism behaves in the following way: $\theta_k \cong 0.9$ if $k \in \text{DUM0}$ and $\theta_k \cong 0.1$ if $k \in \text{DUM1}$, where each response set s_r has a size m of around 500 units. Further, it is not known if unit k belongs to DUM0 or DUM1, in the population U.

In order to estimate *Y*, two calibration models are applied to respondent subsets:

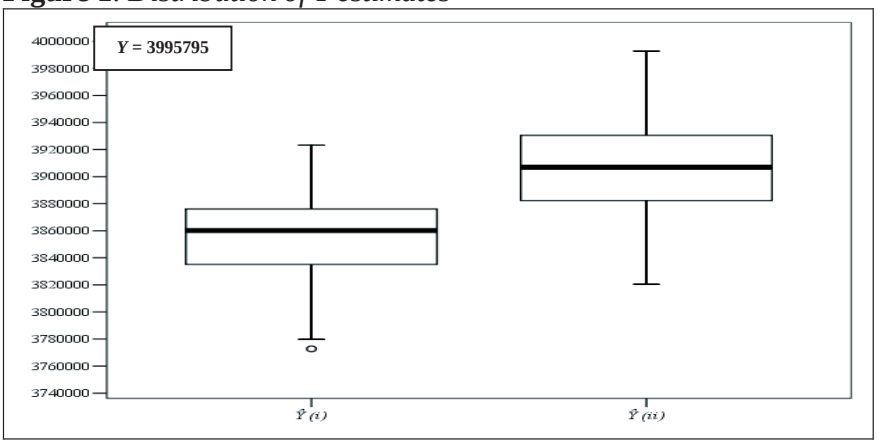
i) *Base* – minimises the distance function, between the initial Horvitz-Thompson weights and the calibrated weights, subject to the constraint that the estimates of auxiliary variables using the calibrated weights will match the known population totals.

ii) *Incorporate variance and covariance structure* – minimises the distance function, between the initial Horvitz-Thompson weights and the calibrated weights, subject to the constraint that the estimates of auxiliary variables using the calibrated weights will match the known population totals and its corresponding variance and covariance structure.

Figure 1 shows box plots for point estimates, where the true population total is used for baseline comparison.

Both calibration models underestimate the true value of *Y*. However, the empirical evidence shows an additional statistical reduction of the nonresponse bias in the *Y* estimate, obtained by applying variance and covariance equality constraints in the calibration procedure.

Figure 1: *Distribution of Y estimates*



3. Final remarks

The results presented in this paper have substantial and methodological implications. If we believe that the respondents are significantly different from the nonresponse units, then explicit modelling of the missing data mechanism is required. On the other hand, if the functional relationship between response probability and auxiliary variables is not easily modelled, then calibration constraints on the known variance and covariance structure of the auxiliary variables can be also used to reduce nonresponse bias in survey estimates.

Before concluding the paper we should point out that the proposed optimal calibration approach requires that each auxiliary variables is known for every unit of the population. Furthermore, caution should be adopted in presence of small samples, since the use of additional auxiliary constraints could cause instability of estimators when the presence of outlying or influential observations is detected. Finally, even if the performance of the proposed calibration estimator has been investigated by means of a simulation study, further empirical investigation is needed when it is applied to small sample size.

References

Deville, J.-C. and Särndal, C.-E. (1992). Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association*, 87, 376-382.

Dupont, F. (1994) Calibration Used as a Nonresponse Adjustment, IN: Diday, E. (ed.) *New Approaches in Classification and Data Analysis*, Springer Verlag, pp.539-548.

Lundström, S. and Särndal, C.-E. (1999). Calibration as a Standard Method for Treatment of Nonresponse. *Journal of Official Statistics*, 15, 305-327.

Särndal, CE, Swensson, B., and Wretman J. (1992). Model Assisted Survey Sampling. New York: Springer-Verlag.