Calibration estimation in dual frame surveys

Maria Giovanna Ranalli and Annalisa Teodoro

Abstract Multiple frame surveys are increasingly used by large statistical agencies and private organizations to reduce frame undercoverage errors and also sampling costs. Estimation for multiple frame surveys has been considered ever since the work of Hartley (1962). In this work we extend the tools of calibration estimation developed so far for single frame surveys to the case of dual frame surveys. Calibration allows to handle different types of auxiliary information and can be shown to encompass as a special case the pseudo empirical maximum likelihood approach recently proposed by Rao and Wu (2010).

Key words: Auxiliary information, Kullback-Leibler distance, Raking ratio.

1 Introduction

To improve estimates, survey statisticians make use of the available auxiliary information either at the design or at the estimation stage. One important example of the latter is given by calibration estimation (Deville and Särndal, 1992), that seeks for new weights that are as close as possible (in terms of a given distance) to the basic design weights and that, at the same time, match benchmark constraints on available auxiliary information (e.g. population totals or means of auxiliary variables). Recently, multiple frame surveys have gained much attention and became largely used by statistical agencies and private organizations to decrease sampling costs – e.g. by using different modes of interviewing in different frames – or to reduce frame undercoverage errors that could occur with the use of only a single sampling frame. Much attention has been devoted in the literature to the introduction of different ways of combining estimates coming from the different frames (see Lohr, 2009, for a recent review). In this work we will extend the calibration paradigm to

Maria Giovanna Ranalli · Annalisa Teodoro
Dept. of Economics, Finance and Statistics University of Perugia, e-mail: giovanna@stat.unipg.it

the estimation of the total of a variable of interest in dual frame surveys as a general tool to include auxiliary information, also available at different levels.

2 Calibration estimation for dual-frame surveys

Let A and B denote two sampling-frames, both can be incomplete, but it is assumed that together cover the entire finite population $\mathscr{U} = \{1,\dots,k,\dots,N\}$. Let $\mathscr{A}$ be the set of population units in frame A and $\mathscr{B}$ the set of population units in frame B. The population of interest, $\mathscr{U}$, may be divided into three mutually exclusive domains, $a = \mathscr{A} \cap \mathscr{B}^c, b = \mathscr{A}^c \cap \mathscr{B}$ and $ab = \mathscr{A} \cap \mathscr{B}$. It is convenient to create a duplicate domain $ba = \mathscr{B} \cap \mathscr{A}$, which is identical to $ab = \mathscr{A} \cap \mathscr{B}$ (see Rao and Wu, 2010). Let N, N_A, N_B, N_a, N_b, N_{ab}, N_{ba} be the number of population units in $\mathscr{U}$, A, B, a, b, ab, ba, respectively. It follows that $N_A = N_a + N_{ab}$, $N_B = N_b + N_{ab}$ and $N = N_a + N_b + N_{ab} = N_a + N_b + N_{ba}$. Let $\delta_k(a) = 1$ if $k \in a$ and 0 otherwise, so that $\sum_{k=1}^{N} \delta_k(a) = N_a$. Let $\delta_k(b)$, $\delta_k(ab)$ and $\delta_k(ba)$ be defined similarly.

The objective is to estimate the finite population total $Y = \sum_{k=1}^{N} y_k$ of a variable of interest y. Note that

$$Y = Y_a + \eta Y_{ab} + (1-\eta) Y_{ba} + Y_b, \tag{1}$$

where $Y_a = \sum_{k=1}^{N} \delta_k(a) y_k$, Y_{ab}, Y_{ba} and Y_b are defined similarly, and $0 \le \eta \le 1$. To this end, two probability samples s_A and s_B are drawn independently from frame A and frame B of sizes n_A and n_B, respectively. Each design induces first-order inclusion probabilities π_k^A and π_k^B, respectively, and sampling weights $d_k^A = 1/\pi_k^A$ and $d_k^B = 1/\pi_k^B$. The sample s_A can be post-stratified as $s_A = s_a \cup s_{ab}$, where $s_a = s_A \cap a$ and $s_{ab} = s_A \cap ab$. Similarly, $s_B = s_b \cup s_{ba}$, where $s_b = s_B \cap b$ and $s_{ba} = s_B \cap (ba)$. Note that s_{ab} and s_{ba} are both from the same domain ab, but s_{ab} is part of the frame A sample and s_{ba} is part of the frame B sample. In this way, we have a sort of "poststratified" sample $s = s_a \cup s_{ab} \cup s_{ba} \cup s_b$ with "poststratum" sample sizes $(n_a, n_{ab}, n_{ba}, n_b)$. Note that $n_A = n_a + n_{ab}$ and $n_B = n_b + n_{ba}$.

If no auxiliary information is available at the estimation stage, each component of (1) can be estimated by its Horvitz-Thompson estimator ($\hat{Y}_a = \sum_{k \in s_A} d_k^A \delta_k(a) y_k$, and similarly for the other components, see e.g. Hartley, 1962). The value of η can be chosen to minimize the variance of $\hat{Y}$. Such a value depends on the variable y except in the case of simple random sampling from both frames (see Lohr, 2009, for a review of alternative methods).

Now, let $\mathbf{x}_k$ be the value taken on unit k by a vector of auxiliary variables of which we assume to know the population total $\mathbf{t}_x = \sum_{k=1}^{N} \mathbf{x}_k$. Using the calibration paradigm (Deville and Särndal, 1992), we wish to modify the aforementioned basic Horvitz-Thompson estimator to obtain a more accurate estimation of the total Y. In particular, let the basic design weights in each post-stratum be $d_{a_k} = d_k^A \delta_k(a)$, $d_{ab_k} = d_k^A \delta_k(ab)$, $d_{ba_k} = d_k^B \delta_k(ba)$ and $d_{b_k} = d_k^B \delta_k(b)$. Then, we wish to find new weights $(w_{a_k}, w_{ab_k}, w_{ba_k}, w_{b_k})$ that are as close as possible to the basic weights, in

terms of the following distance

$$\sum_{k\in s_a} G(w_{a_k},d_{a_k})+\eta\sum_{k\in s_{ab}} G(w_{ab_k},d_{ab_k})+(1-\eta)\sum_{k\in s_{ba}} G(w_{ba_k},d_{ba_k})+\sum_{k\in s_b} G(w_{b_k},d_{b_k}) \tag{2}$$

and satisfy benchmark constraints on the known population totals. Let the distance function in (2) satisfy the usual conditions required in calibration estimation (see e.g. the set of distance functions provided in Deville and Särndal, 1992). We will now consider three simple examples of the vector $\mathbf{x}$.

Case 1. N_A**,** N_B **and** N_{ab} **are all known**.

Let $\mathbf{x}_k=(\delta_k(a),\delta_k(ab),\delta_k(ba),\delta_k(b))$, for $k=1,\ldots,N$. Then, $\mathbf{t}_x=(N_a,N_{ab},N_{ba},N_b)$ and the four calibration constraints can be written as

$$\sum_{k\in s_a} w_{a_k}=N_a,\quad \sum_{k\in s_{ab}} w_{ab_k}=N_{ab},\quad \sum_{k\in s_{ba}} w_{ba_k}=N_{ba},\quad \sum_{k\in S_b} w_{b_k}=N_b. \tag{3}$$

It can be shown that regardless of the choice of the distance measure $G(\cdot,\cdot)$, the new weights will take the Hajek form

$$w_{a_k}=d_{a_k}\frac{N_a}{\hat{N}_a},\quad w_{ab_k}=d_{ab_k}\frac{N_{ab}}{\hat{N}_{ab}},\quad w_{ba_k}=d_{ba_k}\frac{N_{ba}}{\hat{N}_{ba}},\quad w_{b_k}=d_{b_k}\frac{N_b}{\hat{N}_b}. \tag{4}$$

Note that if we take as the distance function in (2) the Kullback-Leibler divergence, which is defined for the first term as $G(w_{a_k},d_{a_k})=\sum_{k\in s_a} d_{a_k}\log(d_{a_k}/w_{a_k})=\sum_{k\in s_a} d_{a_k}\log(d_{a_k})-\sum_{k\in s_a} d_{a_k}\log(w_{a_k})$, i.e. the case 4 distance in Deville and Särndal (1992), we can simply verify that minimizing this distance is equivalent to maximizing the second member on the right, that is equivalent to the maximum PEL method proposed by Rao and Wu (2010).

Case 2. N_A**,** N_B **are known and** N_{ab} **is unknown**.

We can consider this as a case of *incomplete post-stratification* (see e.g. Deville et al., 1993), of which raking ratio is a particular case. In this case $\mathbf{x}_k=(\delta_k(a)+\delta_k(ab),\delta_k(ba)+\delta_k(b))$, for $k=1,\ldots,N$. Then, $\mathbf{t}_x=(N_A,N_B)$ and the two calibration constraints are given by $\sum_{k\in s_a} w_{a_k}+\sum_{k\in s_{ab}} w_{ab_k}=N_A$, $\sum_{k\in s_{ba}} w_{ba_k}+\sum_{k\in S_b} w_{b_k}=N_B$. If we consider the Euclidean distance in (2), given for the first term by $G(w_{a_k},d_{a_k})=\sum_{k\in s_a}(w_{a_k}-d_{a_k})^2/2d_{a_k}$, then the calibrated weights are

$$w_{a_k}=d_{a_k}N_A/(\hat{N}_a+\hat{N}_{ab})\quad\text{and}\quad w_{ab_k}=d_{ab_k}N_A/(\hat{N}_a+\hat{N}_{ab}) \tag{5}$$

and similarly for w_{ba_k} and w_{b_k}. Since $\sum_{k\in s_a} w_{a_k}=\hat{N}_A^w=\hat{N}_aN_A/(\hat{N}_a+\hat{N}_{ab})$, then we have $w_{a_k}=d_{a_k}\hat{N}_a^w/\hat{N}_a$, that is similar to w_{a_k} in (4) except that N_a is estimated.

Case 3. N_A**,** N_B**,** N_{ab} **and the population total in frame** A **of a variable** x **are known**.

Let $\mathbf{x}_k=(\delta_k(a),\delta_k(ab),\delta_k(ba),\delta_k(b),\delta_k(a)x_k+\delta_k(ab)x_k)$, for $k=1,\ldots,N$. Then, $\mathbf{t}_x=(N_a,N_{ab},N_{ba},N_b,X_A)$, where X_A is the population total of x in frame A, and the calibration constraints can be written as

$$\sum_{k\in s_a} w_{a_k}=N_a,\quad \sum_{k\in s_{ab}} w_{ab_k}=N_{ab},\quad \sum_{k\in s_{ba}} w_{ba_k}=N_{ba},\quad \sum_{k\in S_b} w_{b_k}=N_b,$$

$$\sum_{k\in s_a} w_{a_k} x_k + \eta \sum_{k\in s_{ab}} w_{ab_k} x_k + (1-\eta) \sum_{k\in s_{ba}} w_{ba_k} x_k = X_A. \tag{6}$$

If we consider the Euclidean distance, the calibrated weights for domain s_a are given by

$$w_{a_k} = d_{a_k}\Big[\frac{N_a}{\hat{N}_a} + \lambda\big(\frac{\hat{X}_a}{\hat{N}_a} - x_{a_k}\big)\Big], \tag{7}$$

where λ is a Lagrange multiplier given by $\lambda = (X_A - \hat{X}_A^H)/(\hat{S}^2_{a,x} + \eta\hat{S}^2_{ab,x} + (1-\eta)\hat{S}^2_{ba,x})$ with $\hat{X}_A^H = \hat{X}_a^H + \eta\hat{X}_{ab}^H + (1-\eta)\hat{X}_{ba}^H$, $\hat{S}^2_{a,x} = \sum_{k\in s_a} d_{a_k}(x_{a_k} - \hat{X}_a/\hat{N}_a)^2$ and similarly for $\hat{S}^2_{ab,x}$ and $\hat{S}^2_{ba,x}$. The calibrated weights w_{ab_k} and w_{ba_k} are similar to those in (7) but with quantities referred to the appropriate domain, while weights w_{b_k} are as in (4). When using the weights, the resulting estimator resembles a combined regression estimator. In fact $\hat{Y}_{\text{cal}} = \hat{Y}^H + (X_A - \hat{X}_A^H)\hat{\beta}$ where $\hat{Y}^H$ is the estimate of (1) in which each component is estimated by its Hajek estimator, while

$$\hat{\beta} = \frac{\hat{S}_{a,xy} + \eta\hat{S}_{ab,xy} + (1-\eta)\hat{S}_{ba,xy}}{\hat{S}^2_{a,x} + \eta\hat{S}^2_{ab,x} + (1-\eta)\hat{S}^2_{ba,x}},$$

with $\hat{S}_{a,xy} = \sum_{k\in s_a} d_{a_k}(x_{a_k} - \hat{X}_a/\hat{N}_a)(y_{a_k} - \hat{Y}_a/\hat{N}_a)$ and also for $\hat{S}_{ab,xy}$ and $\hat{S}_{ba,xy}$.

Other types of auxiliary information can of course be considered, as those, for example, in which the auxiliary variable pertains only frame B ($X_B = \sum_{k=1}^N \delta_k(ba) + \delta_k(b)x_k$) or the entire finite population $\mathscr{U}$ ($X = \sum_{k=1}^N x_k$), or a combination of the three. Extension of Case 3 to a vector of auxiliary variables is straightforward. Extension to multiple frame surveys, on the other hand, has to account for different levels of frame membership information (Singh and Mecatti, 2011).

References

Deville, J. and Särndal, C. (1992). Calibration estimators in survey sampling. *J. Am. Stat. Ass.*, pages 376–382.

Deville, J., Särndal, C., and Sautory, O. (1993). Generalized raking procedures in survey sampling. *J. Am. Stat. Ass.*, pages 1013–1020.

Hartley, H. (1962). Multiple frame surveys. In *Proceedings of the Social Statistics Section, Am. Stat. Ass.*, volume 19, page 2.

Lohr, S. (2009). Multiple frame surveys. In *Handbook of Statistics, Sample Surveys: Design, Methods and Applications, (Eds., D. Pfeffermann and C.R. Rao)*, number 29A, pages 71–88. Amsterdam: North Holland.

Rao, J. and Wu, C. (2010). Pseudo–empirical likelihood inference for multiple frame surveys. *J. Am. Stat. Ass.*, 105(492):1494–1503.

Singh, A. and Mecatti, F. (2011). Generalized multiplicity-adjusted horvitz-thompson estimation as a unified approach to multiple frame surveys. *J. Off. Stat.*, 27(4):633–650.