ARTICLE IN PRESS

Journal of Computational and Applied Mathematics ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Journal of Computational and Applied Mathematics

journal homepage: www.elsevier.com/locate/cam

Optimal dimension and optimal auxiliary vector to construct calibration estimators of the distribution function

S. Martínez [a], M. Rueda [b,*], H. Martínez [a], A. Arcos [b]

[a] Department of Mathematics, University of Almería, Spain

[b] Department of Statistics and Operational Research, University of Granada, Spain

ARTICLE INFO

Article history:
Received 20 July 2015
Received in revised form 16 October 2015

MSC:
62D05

Keywords:
Auxiliary information
Calibration technique
Distribution function estimates
Survey sampling

ABSTRACT

The calibration technique (Deville and Särndal, 1992) to estimate the finite distribution function has been studied in several papers. Calibration seeks for new weights close enough to sampling weights according to some distance function and that, at the same time, match benchmark constraints on available auxiliary information. The non smooth character of the finite population distribution function causes certain complexities that are resolved by different authors in different ways. One of these is to have consistency at a number of arbitrarily chosen points. This paper deals with the problem of the optimal selection of the number of points and with the optimal selections of these points, when auxiliary information is used by means of calibration.

1. Introduction

Calibration [1] is the principal theme in many recent articles on estimation in survey sampling [2–7]. Calibration has established itself as an important methodological instrument in large scale production of statistics. Several national statistical agencies have developed software designed to compute weights, usually calibrated to auxiliary information available in administrative registers and other accurate sources.

The calibration approach adapts itself to the estimation of more complex parameters than a population total. Before calibration became popular, several papers considered the estimation of distribution functions, with or without the use of auxiliary information [8–11]. As [12] illustrates, there is more than one way to implement the calibration approach in the estimation of the distribution function. Some applications to missing data problems can be seen in [13,14].

The non smooth character of the finite population distribution function causes certain complexities; these are resolved by different authors in different ways. Furthermore, in some cases it is not possible to find an exact solution of the calibration problem as stated.

The computationally simpler method of [15] is an application of model calibration, in that they calibrate with respect to a population total of predicted y-values. Complete auxiliary information is required. Using the known finite population distribution functions of auxiliary variables, compute first the linear predictions. The calibrated weights are obtained by minimizing the chi-square distance subject to calibration equations stated in terms of the predictions, so as to have consistency at J arbitrarily chosen points. It is suggested that a fairly small number of arbitrarily selected points may suffice.

* Corresponding author.
E-mail address: mrueda@ugr.es (M. Rueda).

http://dx.doi.org/10.1016/j.cam.2016.02.002
0377-0427/

The idea to create many benchmarks based on an auxiliary variable was proposed in [16, Exercise 3.35]. This estimator of median can be shown as a special case of how to use 99 known percentiles of an auxiliary variable.

The question of the optimal values in order to obtain the best estimation under simple random sampling without replacement for an arbitrary number of calibration points can be seen in [17]. This paper shows the optimal size of the chosen points (Section 3) and the optimal vector (Section 4). In Section 5 we define the optimum estimator with estimated optimal vector and in Section 6 we include some numerical comparisons.

2. Calibration estimation of the distribution function

Let $U = \{1, 2, \ldots, N\}$ be a finite survey population from which a realized sample $s = \{1, 2, \ldots, n\}$ is drawn with a measurable design d with first and second order inclusion probabilities π_k and π_{kl}. We note by y_k the main variable and by x_k a vector of auxiliary variables at unit k. The values x_k are known for the entire population but y_k is known only if the kth unit is selected on the sample, s. To estimate the distribution function of the study variable y

$$F_y(t) = \frac{1}{N}\sum_{k\in U} \Delta(t - y_k) \tag{1}$$

where

$$\Delta(t - y_k) = \begin{cases} 1 & \text{if } t \geq y_k \\ 0 & \text{if } t < y_k \end{cases}$$

we consider the calibration approach, which consists in the construction of an estimator $\sum_{k\in s} \omega_k \Delta(t - y_k)$ where the calibration weights ω_k are chosen to minimize their average distance from the basic design weights $d_k = 1/\pi_k$ that are used in the Horvitz–Thompson estimator

$$\widehat{F}_{YHT} = \frac{1}{N}\sum_{k\in s} d_k \Delta(t - y_k) \tag{2}$$

subject to conditions that use the auxiliary information provided by the auxiliary vector x.

The distance measure is most commonly chosen as

$$\Phi_s = \frac{1}{2}\sum_{k\in s} \frac{(\omega_k - d_k)^2}{d_k q_k} \tag{3}$$

where q_k are known positive constants unrelated to d_k. Following [15,18], in the definition of calibration conditions, we consider a pseudo-variable $g_k = \widehat{\beta}' x_k$ for $k = 1, 2, \ldots, N$ where:

$$\widehat{\beta}' = \left(\sum_{k\in s} d_k q_k x_k x_k'\right)^{-1} \cdot \sum_{k\in s} d_k q_k x_k y_k.$$

With the pseudo-variable g, we consider the minimization of (3) subject to the following conditions:

$$\frac{1}{N}\sum_{k\in s} \omega_k \Delta(\mathbf{t_g} - g_k) = F_g(\mathbf{t_g}) \tag{4}$$

with $\mathbf{t_g} = (t_1, \ldots, t_P)'$ is a vector chosen arbitrarily, assuming that

$$t_1 < t_2 < \cdots < t_P.$$

The resulting estimator [15] is given by

$$\widehat{F}_{yc}(t) = \widehat{F}_{YHT}(t) + \left(F_g(\mathbf{t_g}) - \widehat{F}_{GHT}(\mathbf{t_g})\right)' \cdot \widehat{D} \tag{5}$$

where $\widehat{F}_{GHT}$ is the Horvitz–Thompson estimator of F_g and

$$\widehat{D} = T^{-1} \cdot \sum_{k\in s} d_k q_k \Delta(\mathbf{t_g} - g_k)\Delta(t - y_k)$$

with the symmetric matrix T given by:

$$T = \sum_{k\in s} d_k q_k \Delta(\mathbf{t_g} - g_k)\Delta(\mathbf{t_g} - g_k)'$$

The calibration estimator (5) can be obtained if we assume that the inverse of T exists.

Please cite this article in press as: S. Martínez, et al., Optimal dimension and optimal auxiliary vector to construct calibration estimators of the distribution function, Journal of Computational and Applied Mathematics (2016), http://dx.doi.org/10.1016/j.cam.2016.02.002