Article Full-text available

Reduction of optimal calibration dimension with a new optimal auxiliary vector for calibrated estimators of the distribution function

WILEY

Jun 2022 · Mathematical Methods in the Applie... 45(2)

DOI: 10.1002/mma.8431

Lab: M. Rueda's Lab

Sergio Martínez · M. Rueda · María Dolores Illescas

Research Interest Score	1.4
Citations	1
Recommendations	2
Reads ⓘ	25

Learn about stats on ResearchGate

New ✕

The publisher of your research added a full-text

Wiley provided the most up-to-date published version of your research.
This means:

- ⊕ More people can discover and read your work, and it's easier for you to track its impact.
- ⊕ If a publisher full-text was added, any reader with an institutional subscription to Wiley journals can now access your full-text on ResearchGate. Open access full-texts added by Wiley can be discovered and read by anyone.
- ⊖ You can no longer edit the details of the publication page, nor remove a full-text added by a publisher.

You can learn more about the partnership between ResearchGate and the publisher here.

Abstract

The calibration method has been widely used to incorporate auxiliary information in the estimation of various parameters. Specifically, some authors adapted this method to estimate the distribution function, although their proposal is computationally simple, its efficiency depends on the selection of an auxiliary vector of points. This work deals with the problem of selecting the calibration auxiliary vector that minimizes the asymptotic variance of the calibration estimator of distribution function. The optimal dimension of the optimal auxiliary vector is reduced considerably with respect to previous studies so that with a smaller set of points, the minimum of the asymptotic variance can be reached, which in turn allows to improve the efficiency of the estimates.

Share More

WILEY

Wiley added this full-text to your publication. · Learn more

Publicly visible

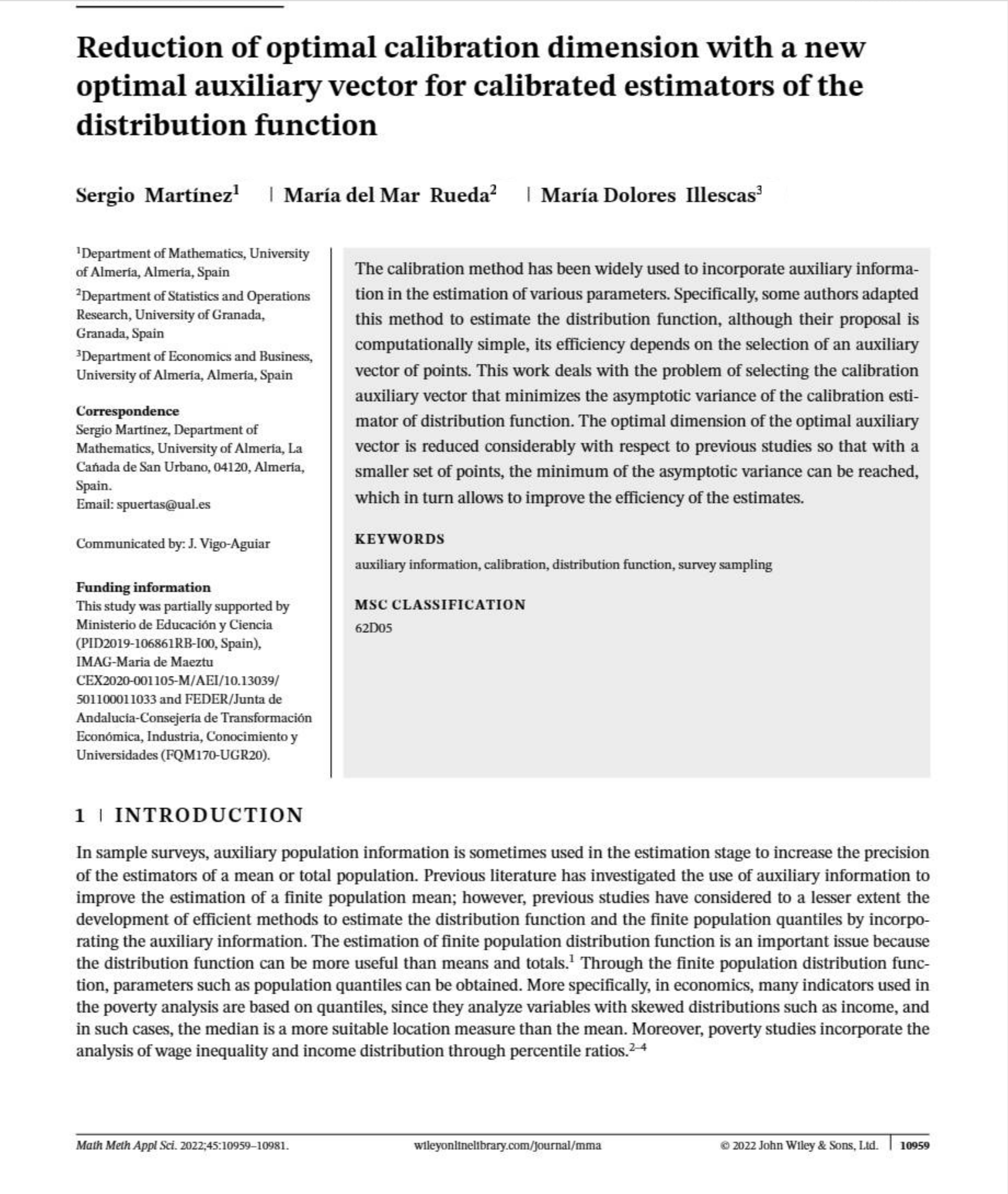
Content available from Mathematical Methods in the Applied Sciences

Reduction of optimal calibration dim... on.pdf

Page 1

Share

More



Reduction of optimal calibration dimension with a new optimal auxiliary vector for calibrated estimators of the distribution function

Sergio Martínez[1] | María del Mar Rueda[2] | María Dolores Illescas[3]

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

[2]Department of Statistics and Operations Research, University of Granada, Granada, Spain

[3]Department of Economics and Business, University of Almería, Almería, Spain

Correspondence
Sergio Martínez, Department of Mathematics, University of Almería, La Cañada de San Urbano, 04120, Almería, Spain.
Email: spuertas@ual.es

Communicated by: J. Vigo-Aguiar

Funding information
This study was partially supported by Ministerio de Educación y Ciencia (PID2019-106861RB-I00, Spain), IMAG-Maria de Maeztu CEX2020-001105-M/AEI/10.13039/501100011033 and FEDER/Junta de Andalucía-Consejería de Transformación Económica, Industria, Conocimiento y Universidades (FQM170-UGR20).

The calibration method has been widely used to incorporate auxiliary information in the estimation of various parameters. Specifically, some authors adapted this method to estimate the distribution function, although their proposal is computationally simple, its efficiency depends on the selection of an auxiliary vector of points. This work deals with the problem of selecting the calibration auxiliary vector that minimizes the asymptotic variance of the calibration estimator of distribution function. The optimal dimension of the optimal auxiliary vector is reduced considerably with respect to previous studies so that with a smaller set of points, the minimum of the asymptotic variance can be reached, which in turn allows to improve the efficiency of the estimates.

KEYWORDS
auxiliary information, calibration, distribution function, survey sampling

MSC CLASSIFICATION
62D05

1 | INTRODUCTION

In sample surveys, auxiliary population information is sometimes used in the estimation stage to increase the precision of the estimators of a mean or total population. Previous literature has investigated the use of auxiliary information to improve the estimation of a finite population mean; however, previous studies have considered to a lesser extent the development of efficient methods to estimate the distribution function and the finite population quantiles by incorporating the auxiliary information. The estimation of finite population distribution function is an important issue because the distribution function can be more useful than means and totals.[1] Through the finite population distribution function, parameters such as population quantiles can be obtained. More specifically, in economics, many indicators used in the poverty analysis are based on quantiles, since they analyze variables with skewed distributions such as income, and in such cases, the median is a more suitable location measure than the mean. Moreover, poverty studies incorporate the analysis of wage inequality and income distribution through percentile ratios.[2–4]

of research in survey sampling.

Previous works[6,11,12] use different implementations of the calibration approach to obtain estimators of the distribution function and the quantiles. Under a general superpopulation model, Wu[13] proposes a model-calibrated estimators that is optimal under a chosen model with respect to the anticipated variance. Although Wu[13] considers a general sampling design, its proposal does not produce an estimator with the properties of a genuine distribution function unless the weight system is obtained by using a point t_0 for any t value, which restricts the efficiency of the estimator to a neighborhood of t_0. Additionally, the proposal[13] requires the estimation of certain superpopulation parameters that depend on the study variable, which may restrict its applicability in some cases and also require additional conditions on the sampling design to maintain the asymptotic behavior of the proposed estimator.[14]

Nonparametric regression[6,15] is also used for model-calibration estimation of the distribution function. Mayor-Gallego et al[16] propose a new estimator for the distribution function that integrates ideas from model calibration and penalized calibration. The method[6] is computationally simple, and it employs the calibration method by minimizing the chi-square distance subject to calibration equations that require the use of arbitrarily fixed values. One drawback of these estimators is that their efficiency depends on selected points. Under simple random sampling, the problem of optimal selection points in order to obtain the best estimation has been treated in previous works.[17–20] In fact, the work[17] obtained the optimal dimension and the optimal auxiliary vector for the estimator of the distribution function proposed in the work,[6] and although this proposal do not generate a unique weight system that is optimal for each point t, it produces an estimator that is computationally simple and is a genuine distribution function that can be used directly in the estimation of quantiles and poverty measures.[21]

In many situations, the optimal auxiliary vector has a very high dimension, which makes the calibration process difficult and can also affect the efficiency of the estimator. Performing calibration with a high-dimensional auxiliary dataset can be several problems: The variance of the calibration estimator can be increased, and the optimization procedure may fail. Nascimento Silva and Skinner[22] showed that if too many auxiliary variables are used, the bias of the calibrated estimator increases and can become nonnegligible compared to the variance (over-calibration). Recently, Chauvet and Goga[23] theoretically prove that overcalibration may deteriorate the efficiency of the estimates. Various procedures have been suggested for variable selection. Nascimento Silva and Skinner[22] computed the mean squared error (MSE) for all possible subsets of quantitative auxiliary variables and then chose the one producing the smallest MSE. Later, Chambers and Clark[24] used forward and stepwise selection based on the difference between the MSE of the prediction for two nested sets of variables. Alternatively, the least absolute shrinkage and selection operator (LASSO)[25] might be considered for selecting the best subsets. Once the best set of regressors has been selected, the calibration is performed on these variables alone. Another approach to consider is that of penalized calibration,[26] which takes account of auxiliary information by attaching more or less importance according to its presumed explanatory power for the variable of interest. In a different way, Cardot et al[27] and Rota[28] suggested applying principal component analysis for quantitative auxiliary variables in order to achieve a strong dimension reduction. These works are oriented to the estimation of linear parameters.

In this work, we intend to analyze whether it is possible to reduce the optimal dimension of the auxiliary vector proposed in the previous work.[17] The remainder of the article is organized as follow. After introducing the problem of distribution function estimation in Section 2 with the method proposed in research work[6] and the optimal auxiliary vector proposed in the previous work,[17] in Section 3, we will analyze the conditions under which we can reduce the dimension of the optimal auxiliary vector. Then, Section 4 proposes a new calibration estimator based on the results of Section 3. Section 5 reports the results of an extensive simulation study run on a set of synthetic and real finite populations in which the performance of the proposed class of estimators is investigated for finite size samples. Section 6 provides some conclusions.

2 | CALIBRATION ESTIMATION OF THE DISTRIBUTION FUNCTION AND OPTIMAL AUXILIARY VECTOR

Let $U = \{1, \dots, N\}$ a finite population composed of N different units, and let $s = \{1, 2, \dots, n\}$ a random sample of size n selected using a specified sampling design $p(\cdot)$ with first- and second-order inclusion probabilities $\pi_k > 0$ and $\pi_{kl} > 0$ $k, l \in U$ respectively, and $d_k = \pi_k^{-1}$ denotes the sampling design-basic weight for unit $k \in U$. Let y_k be the study variable and $\mathbf{x}'_k = (x_{1k}, \dots, x_{Jk})$ be a vector of auxiliary variables at unit k. We assume that value $\mathbf{x}_k$ is available for all population

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

with

$$\Delta(t - y_k) = \begin{cases} 1 \text{ si } t \geq y_k, \\ 0 \text{ si } t < y_k. \end{cases}$$

A design-based estimator of the distribution function $F_y(t)$ is the Horvitz–Thompson estimator, defined by

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

The estimator $\hat{F}_{YHT}(t)$ is unbiased, but it does not incorporate the auxiliary information provided by the auxiliary vector $\mathbf{x}$.

Several authors[6,12,29,30] have incorporated the auxiliary information to obtain new estimators of $F_y(t)$ through the calibration method.[5] The proposal[6] applies the calibration procedure from a pseudo-variable

$$g_k = (\hat{\beta})' \mathbf{x}_k \ \text{ for } \ k = 1, 2, \dots N, \tag{3}$$

$$\hat{\beta} = \left(\sum_{k \in s} d_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \cdot \sum_{k \in s} d_k \mathbf{x}_k y_k. \tag{4}$$

With the variable g, the basic weights d_k are replaced by new calibrated weights ω_k through the minimization of the chi-square distance measure

$$\Phi_s = \sum_{k \in s} \frac{(\omega_k - d_k)^2}{d_k q_k}, \tag{5}$$

subject to the calibration constrains

$$\frac{1}{N} \sum_{k \in s} \omega_k \Delta(t_j - g_k) = F_g(t_j) \ j = 1, 2, \dots, P, \tag{6}$$

where $F_g(t_j)$ denotes the finite distribution function of the pseudo-variable g_k evaluated at the points $t_j,\ j = 1, 2, \dots, P$. We assume, with no loss in generality, $t_1 < t_2 < \dots t_P$. The values q_k are known positive constants unrelated to d_k.

Following Rueda et al,[6] we assume that the matrix T given by:

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

is nonsingular. With this calibration procedure, the calibration estimator obtained is as follows:

$$\hat{F}_{yc}(t) = \hat{F}_{YHT}(t) + \left(F_g(\mathbf{t_g}) - \hat{F}_{GHT}(\mathbf{t_g}) \right)' \cdot \hat{D}(\mathbf{t_g}), \tag{7}$$

where

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

and $\hat{F}_{GHT}(\mathbf{t_g})$ is the Horvitz–Thompson estimator of $F_g(\mathbf{t_g})$ evaluated at $\mathbf{t_g} = (t_1, \dots, t_P)'$.

The calibration estimator $\hat{F}_{yc}(t)$ has the following asymptotic variance[6]:

$$AV(\hat{F}_{yc}(t)) = \frac{1}{N^2} \sum_{k \in U} \sum_{l \in U} \Delta_{kl} (d_k E_k)(d_l E_l) \tag{8}$$

$$D(\mathbf{t_g}) = \left(\sum_{k\in U} q_k \Delta(\mathbf{t_g} - g_k)\Delta(\mathbf{t_g} - g_k)' \right) \cdot \left(\sum_{k\in U} q_k \Delta(\mathbf{t_g} - g_k)\Delta(t - y_k) \right). \tag{9}$$

As a consequence, the behavior of the estimator $\hat{F}_{yc}(t)$ and its precision depends on the selection of the vector $\mathbf{t_g}$.

Previous works[17,19,20] treated, under simple random sampling without replacement and $q_k = c$ for all $k \in U$, the optimal selection of the vector $\mathbf{t_g}$ in order to minimize the asymptotic variance (8). In fact, Martínez et al[17] established the optimal dimension of $\mathbf{t_g}$ and its optimal value, for a given value t, through the definition of the sets:

$$A_t = \{g_k : k \in U;\ y_k \leq t\} = \{a_1^t, a_2^t, \dots, a_{M_t}^t\} \text{ with } a_h^t < a_{h+1}^t \text{ for } h = 1, \dots, M_t - 1, \tag{10}$$

where M_t is the number of elements in the set A_t and

$$B_t = \{b_1^t, b_2^t, \dots, b_{M_t}^t\} \tag{11}$$

with

$$b_1^t = \max_{l\in U_1}\{g_l\} \text{ where } U_1 = \{l \in U : g_l < a_1^t\},$$
$$b_h^t = \max_{l\in U_h}\{g_l\} \text{ where } U_h = \{l \in U : a_{h-1}^t < g_l < a_h^t\}\ h = 2, 3, \dots, M_t,$$

and $b_h^t < b_{h+1}^t$ for $h = 1, \dots, M_t - 1$.

Thus, Martínez et al[17] established that the auxiliary vector $\mathbf{t}_g$ has optimal dimension $P = 2M_t$ if b_h^t exists for $h = 1, \dots, M_t$ and the optimal value of $\mathbf{t}_g$ is given by

$$\mathbf{t_{OPT}}(t) = \left(b_1^t, a_1^t, \dots, b_{M_t}^t, a_{M_t}^t\right). \tag{12}$$

If there are some values $j_1^t, j_2^t, \dots j_{p_t}^t \in \{1, \dots, M_t\}$; such as $b_{j_h}^t$ does not exits for $h = 1, 2, \dots p_t$ with $p_t \leq M_t$ and $j_h^t \neq j_q^t$ if $h \neq q$, the optimal dimension is given by $P = 2M_t - p_t$ and the optimal auxiliary vector $\mathbf{t_{OP}}$ is as follows:

$$\mathbf{t_{OP}}(t) = (b_1^t, a_1^t \dots, b_{j_1-1}^t, a_{j_1-1}^t, a_{j_1}^t, b_{j_1+1}^t, \dots, b_{j_h-1}^t, a_{j_h-1}^t, a_{j_h}^t, b_{j_h+1}^t, \dots b_{M_t}^t, a_{M_t}^t). \tag{13}$$

In the next section, we will analyze if the minimum of the asymptotic variance can be reached with a vector of less dimension, and we will establish conditions under which the dimension of the optimal vector $\mathbf{t_{OPT}}(t)$ can be reduced under simple random sampling without replacement.

3 | DIMENSION REDUCTION OF THE OPTIMAL AUXILIARY VECTOR

In this section, we will analyze the conditions under which the dimension of the optimal vector $\mathbf{t_{OPT}}(t)$ can be reduced; that is, we will analyze the existence of a vector with a smaller dimension than $\mathbf{t_{OPT}}(t)$ that allows obtaining the minimum value of the asymptotic variance of the estimator $\hat{F}_{yc}(t)$.

For the minimization of the asymptotic variance (8), we consider it as a function of a vector $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_P)$ of dimension P:

$$AV(\hat{F}_{yc}(t)) = \frac{1}{N^2} \sum_{k\in U} \sum_{l\in U} \Delta_{kl}(d_k\Gamma_k)(d_l\Gamma_l), \tag{14}$$

with $\Gamma_k = \Delta(t - y_k) - \Delta(\boldsymbol{\gamma} - g_k) \cdot D(\boldsymbol{\gamma})$, with $D(\boldsymbol{\gamma})$ given by (9).

Following Martínez et al,[20] under simple random sampling without replacement and $q_k = c$ for all units in the population, the minimization of (14) is equivalent to the minimization of the function:

$$Q_t(\boldsymbol{\gamma}) = Q_t(\gamma_1, \dots, \gamma_P) = 2NF_y(t) \cdot K_t(\gamma_P) - \sum_{j=1}^{P} \frac{\left(K_t(\gamma_j) - K_t(\gamma_{j-1})\right)^2}{(F_g(\gamma_j) - F_g(\gamma_{j-1}))} - (K_t(\gamma_P))^2, \tag{15}$$

with $K_t(\gamma_j) = \sum_{k\in U} \Delta(\gamma_j - g_k)\Delta(t - y_k)$, where we suppose that $F_g(\gamma_0) = 0$ and $K_t(\gamma_0) = 0$.

a plenty of constraints which raises the computational cost for calculating the estimator. For example, if we consider $t = y_{max}$ where

$$y_{max} = \max_{k \in U} y_k,$$

the optimal auxiliary vector $\mathbf{t_{OP}}(t) = (a_1, a_2, \dots, a_M)$ can be reduced to the auxiliary vector $\boldsymbol{\gamma} = (a_M)$ (see Appendix A.1.1). Consequently, the optimal dimension can be reducted from M to 1.

In a similar way, we try to reduce the dimension of the auxiliary vector to reach the minimum of $Q_t(\boldsymbol{\gamma})$. For it, given a value t for which we want to estimate $F_y(t)$, we consider the sets A_M, A_t and B_t given by (A1); (10) and (11) respectively and for each $a_i \in A_M$, we define:

$$r_i = \text{Frequency of the } \ a_i.$$

For the value t, we have:

$$A_t = \left\{a_1^t, a_2^t, \dots, a_{M_t}^t\right\} = \left\{a_{f_1^t}, a_{f_2^t}, \dots, a_{f_{M_t}^t}\right\},$$

where

$$\left\{f_1^t, f_2^t, \dots, f_{M_t}^t\right\} \subseteq \{1, 2, \dots, M\} \ \text{ and } \ f_1^t < f_2^t < \cdots < f_{M_t}^t.$$

Similarly, we consider the following set:

$$C_t = \left\{g_k \, : \, k \in U; \ y_k > t\right\} = \left\{c_1^t, c_2^t, \dots, c_{S_t}^t\right\} = \left\{a_{l_1^t}, a_{l_2^t}, \dots, a_{l_{S_t}^t}\right\},$$

with

$$\left\{l_1^t, l_2^t, \dots, l_{S_t}^t\right\} \subseteq \{1, 2, \dots, M\} \ \text{ and } \ l_1^t < l_2^t < \cdots < l_{S_t}^t.$$

It is clear that $A_t \cup C_t = A_M$ and since for two different units k and j can be possible that $g_j = g_k = a_i$ and $y_k > t$ and $y_j < t$, not necessarily $A_t \cap C_t = \emptyset$. For the sets A_t and C_t, we define:

$$p_i^t = \text{Frequency of the } \ a_i^t \ \text{ in } \ A_t,$$
$$q_i^t = \text{Frequency of the } \ c_i^t \ \text{ in } \ C_t.$$

Next, we consider the following sets:

$$D_t = \{c_i \in C_t \, : \, q_i^t = r_i\}, \tag{16}$$

$$Z_t = \{a_i^t \in A_t \, : \, q_i^t = 0\} = \{a_i^t \in A_t \, : \, a_i^t \notin C_t\} = A_t - C_t, \tag{17}$$

$$F_t = \{a_i^t \in A_t \, : \, 0 < q_i^t < r_i\}. \tag{18}$$

It is easy to see that $D_t = A_M - A_t$ and consequently $A_t \cap D_t = \emptyset$. Furthermore, $B_t \subseteq D_t$; $A_t = Z_t \cup F_t$ and $Z_t \cap F_t = \emptyset$.

Firstly, if we suppose that $D_t = A_M$, we have $A_t = \emptyset$ and consequently $y_k > t$, $\forall k \in U$. In this case, $F_y(t) = 0$, and we can calibrate with any auxiliary vector since

$$\hat{F}_{yc}(t) = \frac{1}{N} \sum_{k \in s} \omega_k \Delta(t - y_k) = 0,$$

regardless of the auxiliary vector, so we can calibrate with $\mathbf{t_{OP}}(t) = a_M$ with Dimension 1.

Secondly, if we suppose that $D_t = \emptyset$, then $B_t = \emptyset$ and $A_t = A_M$. In this case, following Martínez et al,[17] the optimal auxiliary vector $\mathbf{t_{OP}}(t) = (a_1, a_2, \dots, a_M)$.

Since $A_t = Z_t \cap F_t = A_M$, if we suppose that $Z_t = A_t = A_M$, then $t > y_k$ $\forall k \in U$, and this case is like the case where $t = y_{max}$ and although the optimal auxiliary vector is $\mathbf{t_{OP}}(t) = (a_1, a_2, \dots, a_M)$, we can reach the minimum value of $Q_\gamma(t)$ with the auxiliary vector $\gamma = (a_M)$.

On the other hand, if we consider that $Z_t = \emptyset$ and $F_t = A_M$, there is not reduction in the optimal auxiliary vector $\mathbf{t_{OP}}(t)$ (see Appendix A.2).

Next, if we suppose that $Z_t \neq A_t = A_M$ and $F_t \neq A_t = A_M$, then there is a set $I_{F_t} = \{j_1, j_2, \dots, j_l\} \subseteq \{1, 2, \dots M\}$ such that $a_{j_i} \in F_t$ and therefore $q_{j_i}^t \neq 0$ for $i = 1, 2, \dots, l$.

$$
\begin{aligned}
K_t(a_1) &= \sum_{k\in U} \Delta(a_1 - g_k)\Delta(t - y_k) = NF_g(a_1) \\
\vdots \qquad & \qquad\qquad\qquad \vdots \\
K_t(a_{(j_1-1)}) &= \sum_{k\in U} \Delta(a_{j_1-1} - g_k)\Delta(t - y_k) = NF_g(a_{j_1-1}).
\end{aligned}
$$

Similarly, for $j_i, \ldots j_{i+1} - 1$ with $i = 1, 2, \ldots l - 1$, we have:

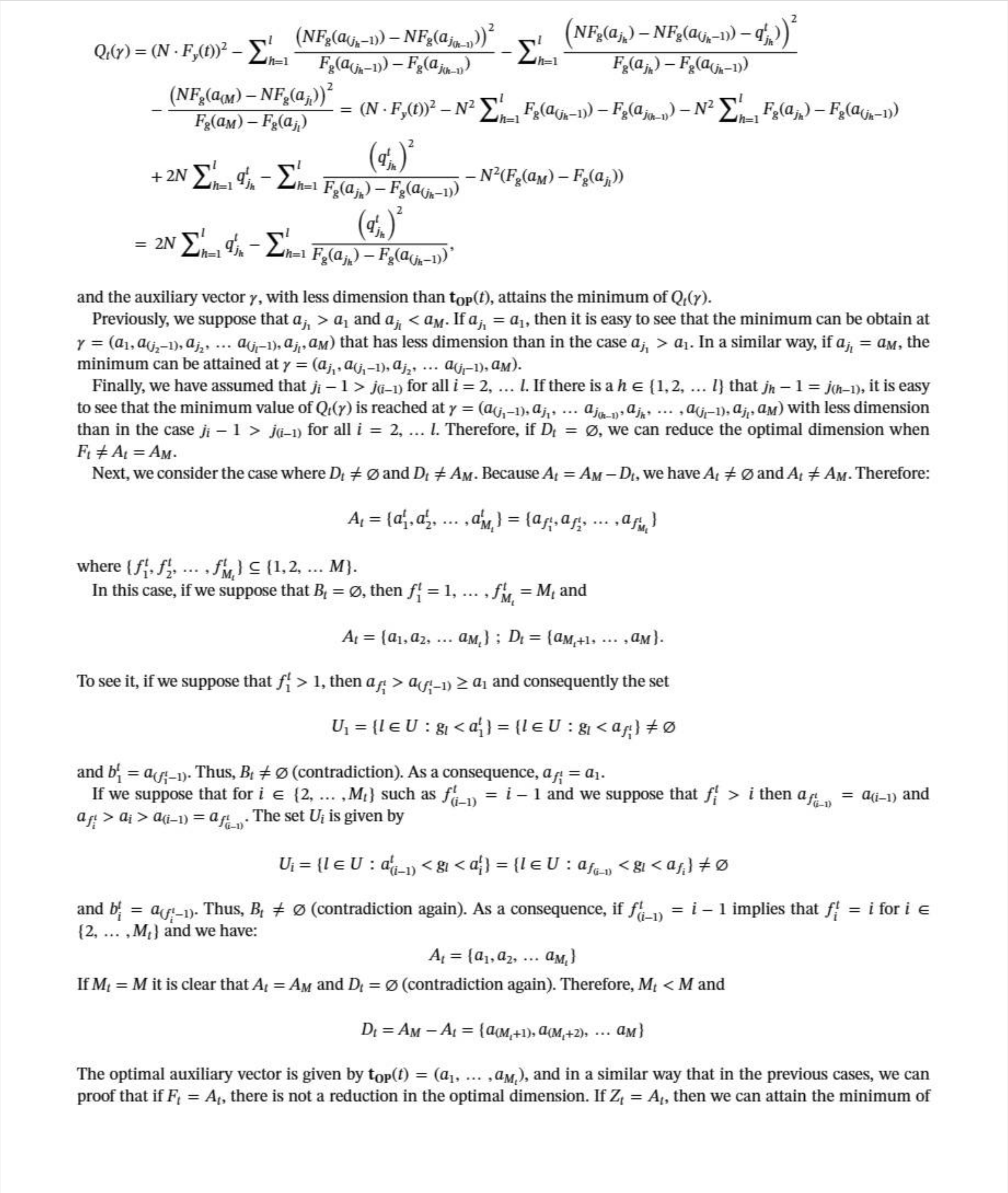
$$
\begin{aligned}
K_t(a_{j_i}) &= \sum_{k\in U} \Delta(a_{j_i} - g_k)\Delta(t - y_k) = NF_g(a_{j_i}) - \sum_{h=1}^{i} q^t_{j_h} \\
\vdots \qquad & \qquad\qquad\qquad \vdots \\
K_t(a_{(j_{(i+1)}-1)}) &= \sum_{k\in U} \Delta(a_{(j_{(i+1)}-1)} - g_k)\Delta(t - y_k) = NF_g(a_{(j_{(i+1)}-1)}) - \sum_{h=1}^{i} q^t_{j_h}
\end{aligned}
$$

and finally, for $j_l, \ldots, M$

$$
\begin{aligned}
K_t(a_{j_l}) &= \sum_{k\in U} \Delta(a_{j_l} - g_k)\Delta(t - y_k) = NF_g(a_{j_l}) - \sum_{h=1}^{l} q^t_{j_h} \\
\vdots \qquad & \qquad\qquad\qquad \vdots \\
K_t(a_M) &= \sum_{k\in U} \Delta(a_M - g_k)\Delta(t - y_k) = NF_g(a_M) - \sum_{h=1}^{l} q^t_{j_h} = NF_y(t).
\end{aligned}
$$

The minimum of $Q_t(\gamma)$ reached at the optimum auxiliary vector $\mathbf{t_{OP}}(t)$ is given by

$$
\begin{aligned}
Q_t(\mathbf{t_{OP}}(t)) &= (NF_y(t))^2 - \sum_{j=1}^{M} \frac{\left(K_t(a_j - K_t(a_{j-1}))\right)^2}{F_g(a_j) - F_g(a_{j-1})} = \\
&= (NF_y(t))^2 - N^2 \cdot \sum_{\substack{j=1 \\ j\notin\{j_1,\ldots,j_l\}}}^{M} \frac{\left(F_g(a_j) - F_g(a_{j-1})\right)^2}{F_g(a_j) - F_g(a_{j-1})} - \sum_{j\in\{j_1,\ldots,j_l\}} \frac{\left(NF_g(a_j) - NF_g(a_{j-1})) - q^t_j\right)^2}{F_g(a_j) - F_g(a_{j-1})} \\
&= (NF_y(t))^2 - N^2 \cdot \sum_{\substack{j=1 \\ j\notin I_{F_t}}}^{M} \left(F_g(a_j) - F_g(a_{j-1}))\right) - N^2 \cdot \sum_{j\in I_{F_t}} \left(F_g(a_j) - F_g(a_{j-1})\right) \\
&\quad + 2N \sum_{j\in I_{F_t}} q^t_j - \sum_{j\in I_{F_t}} \frac{\left(q^t_j)\right)^2}{F_g(a_j) - F_g(a_{j-1})} = \\
&= (NF_y(t))^2 - N^2 \cdot \sum_{j=1}^{M} \left(F_g(a_j) - F_g(a_{j-1}))\right) + 2N \sum_{j\in I_{F_t}} q^t_j - \sum_{j\in I_{F_t}} \frac{\left(q^t_j\right)^2}{F_g(a_j) - F_g(a_{j-1})} = \\
&= (NF_y(t))^2 - N^2 + 2N \sum_{j\in I_{F_t}} q^t_j - \sum_{j\in I_{F_t}} \frac{\left(q^t_j\right)^2}{F_g(a_j) - F_g(a_{j-1})}.
\end{aligned} \tag{19}
$$



$$Q_t(\gamma) = (N \cdot F_y(t))^2 - \sum\nolimits_{h=1}^{l} \frac{\left(NF_g(a_{(j_h-1)}) - NF_g(a_{j_{(h-1)}})\right)^2}{F_g(a_{(j_h-1)}) - F_g(a_{j_{(h-1)}})} - \sum\nolimits_{h=1}^{l} \frac{\left(NF_g(a_{j_h}) - NF_g(a_{(j_h-1)}) - q^t_{j_h})\right)^2}{F_g(a_{j_h}) - F_g(a_{(j_h-1)})}$$

$$- \frac{\left(NF_g(a_{(M)}) - NF_g(a_{j_l})\right)^2}{F_g(a_M) - F_g(a_{j_l})} = (N \cdot F_y(t))^2 - N^2 \sum\nolimits_{h=1}^{l} F_g(a_{(j_h-1)}) - F_g(a_{j_{(h-1)}}) - N^2 \sum\nolimits_{h=1}^{l} F_g(a_{j_h}) - F_g(a_{(j_h-1)})$$

$$+ 2N \sum\nolimits_{h=1}^{l} q^t_{j_h} - \sum\nolimits_{h=1}^{l} \frac{\left(q^t_{j_h}\right)^2}{F_g(a_{j_h}) - F_g(a_{(j_h-1)})} - N^2(F_g(a_M) - F_g(a_{j_l}))$$

$$= 2N \sum\nolimits_{h=1}^{l} q^t_{j_h} - \sum\nolimits_{h=1}^{l} \frac{\left(q^t_{j_h}\right)^2}{F_g(a_{j_h}) - F_g(a_{(j_h-1)})},$$

and the auxiliary vector γ, with less dimension than $\mathbf{t_{OP}}(t)$, attains the minimum of $Q_t(\gamma)$.

Previously, we suppose that $a_{j_1} > a_1$ and $a_{j_l} < a_M$. If $a_{j_1} = a_1$, then it is easy to see that the minimum can be obtain at $\gamma = (a_1, a_{(j_2-1)}, a_{j_2}, \dots a_{(j_l-1)}, a_{j_l}, a_M)$ that has less dimension than in the case $a_{j_1} > a_1$. In a similar way, if $a_{j_l} = a_M$, the minimum can be attained at $\gamma = (a_{j_1}, a_{(j_1-1)}, a_{j_2}, \dots a_{(j_l-1)}, a_M)$.

Finally, we have assumed that $j_i - 1 > j_{(i-1)}$ for all $i = 2, \dots l$. If there is a $h \in \{1, 2, \dots l\}$ that $j_h - 1 = j_{(h-1)}$, it is easy to see that the minimum value of $Q_t(\gamma)$ is reached at $\gamma = (a_{(j_1-1)}, a_{j_1}, \dots a_{j_{(h-1)}}, a_{j_h}, \dots, a_{(j_l-1)}, a_{j_l}, a_M)$ with less dimension than in the case $j_i - 1 > j_{(i-1)}$ for all $i = 2, \dots l$. Therefore, if $D_t = \varnothing$, we can reduce the optimal dimension when $F_t \neq A_t = A_M$.

Next, we consider the case where $D_t \neq \varnothing$ and $D_t \neq A_M$. Because $A_t = A_M - D_t$, we have $A_t \neq \varnothing$ and $A_t \neq A_M$. Therefore:

$$A_t = \{a^t_1, a^t_2, \dots, a^t_{M_t}\} = \{a_{f^t_1}, a_{f^t_2}, \dots, a_{f^t_{M_t}}\}$$

where $\{f^t_1, f^t_2, \dots, f^t_{M_t}\} \subseteq \{1, 2, \dots M\}$.

In this case, if we suppose that $B_t = \varnothing$, then $f^t_1 = 1, \dots, f^t_{M_t} = M_t$ and

$$A_t = \{a_1, a_2, \dots a_{M_t}\}\ ;\ D_t = \{a_{M_t+1}, \dots, a_M\}.$$

To see it, if we suppose that $f^t_1 > 1$, then $a_{f^t_1} > a_{(f^t_1-1)} \geq a_1$ and consequently the set

$$U_1 = \{l \in U : g_l < a^t_1\} = \{l \in U : g_l < a_{f^t_1}\} \neq \varnothing$$

and $b^t_1 = a_{(f^t_1-1)}$. Thus, $B_t \neq \varnothing$ (contradiction). As a consequence, $a_{f^t_1} = a_1$.

If we suppose that for $i \in \{2, \dots, M_t\}$ such as $f^t_{(i-1)} = i - 1$ and we suppose that $f^t_i > i$ then $a_{f^t_{(i-1)}} = a_{(i-1)}$ and $a_{f^t_i} > a_i > a_{(i-1)} = a_{f^t_{(i-1)}}$. The set U_i is given by

$$U_i = \{l \in U : a^t_{(i-1)} < g_l < a^t_i\} = \{l \in U : a_{f_{(i-1)}} < g_l < a_{f_i}\} \neq \varnothing$$

and $b^t_i = a_{(f^t_i-1)}$. Thus, $B_t \neq \varnothing$ (contradiction again). As a consequence, if $f^t_{(i-1)} = i - 1$ implies that $f^t_i = i$ for $i \in \{2, \dots, M_t\}$ and we have:

$$A_t = \{a_1, a_2, \dots a_{M_t}\}$$

If $M_t = M$ it is clear that $A_t = A_M$ and $D_t = \varnothing$ (contradiction again). Therefore, $M_t < M$ and

$$D_t = A_M - A_t = \{a_{(M_t+1)}, a_{(M_t+2)}, \dots a_M\}$$

The optimal auxiliary vector is given by $\mathbf{t_{OP}}(t) = (a_1, \dots, a_{M_t})$, and in a similar way that in the previous cases, we can proof that if $F_t = A_t$, there is not a reduction in the optimal dimension. If $Z_t = A_t$, then we can attain the minimum of when the auxiliary information includes qualitative variables, traditional calibration techniques may be not feasible or the optimisation...

14 Reads · 1 Citation

Comments on: Deville and Särndal's calibration: revisiting a 25 years old successful optimization problem

Article Private full-text

November 2019 · Test

M. Rueda

40 Reads · 3 Citations

Methods to Counter Self-Selection Bias in Estimations of the Distribution Function and Quantiles

Article Full-text available

December 2022

M. Rueda · Sergio Martínez-Puertas · Luis Castro-Martín

Many surveys are performed using non-probability methods such as web surveys, social networks surveys, or opt-in panels. The estimates made from these data sources are usually biased and must be adjusted to make them representative of the target population. Techniques to mitigate this selection bias in non-probability samples often involve calibration, propensity score adjustment, or statistical matching. In...

42 Reads

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

Article Private full-text

February 2016 · Journal of Computational and Applied Mathematics

S. Martínez · M. Rueda · Honorina Martínez · A. Arcos

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...

80 Reads · 8 Citations

View more related research