Revista Colombiana de Estadística
Junio 2009, volumen 32, no. 1, pp. 79 a 97

Estimation Stage in Survey Sampling: A Multiparameter Approach

Estimación en encuestas por muestreo: un enfoque multiparamétrico

Hugo Andrés Gutiérrez[a]

Centro de Investigaciones y Estudios Estadísticos (CIEES), Facultad de Estadística, Universidad Santo Tomás, Bogotá, Colombia

Abstract

Most of the real applications in survey sampling involve not one, but several characteristics of study. In this paper, a unified framework of joint estimation of the parameters of interest is presented under various sampling designs. The applications of the results of this research entails a significant gain in computational efficiency.

***Key words*:** Multipurpose survey, Multiparameter estimation.

Resumen

La mayoría de las aplicaciones en encuestas por muestreo involucran múltiples variables de estudio. En este artículo se presenta un marco de referencia para la estimación conjunta de los parámetros de interés en algunos diseños de muestreo. La aplicación de los resultados encontrados garantizan una ganancia significativa en la eficiencia computacional.

***Palabras clave*:** diseño de encuestas complejas, estimación de parámetros.

1. Introduction

The purpose of sample surveys is to gather information about a certain finite population by estimating parameters such as means, totals, or ratios. However, most surveys do not include just one but several study variables. Sampling books seem to discard the fact that a survey is seldom interested in the estimation of one single parameter and most of the theory developed by sampling researchers is focused on the search of both, design and estimator, that perform well and gain in efficiency under the single parameter case. Several advantages have been developed in this field. However, all of them are motivated by the assumption

[a]Director. E-mail: hugogutierrez@usantotomas.edu.co

that the survey methodologist is interested in one single parameter. As Holmberg (2002*a*) claims, «a typical business survey has several study variables and several target parameters... with multiple target parameters, and multiple requirements on precision, the practising statistician must then select a compromise design.»

A survey could be divided in two stages: The design stage and the estimation stage. The work of Holmberg during the past decade was related to find a sampling design that yields unequal first order inclusion probabilities, and that is optimal in the sense that significant improvements on the overall precision are possible. This paper is focused on the next step after planning the survey, the estimation stage, presenting a methodological unified framework in the estimation of each parameter of interest by means of the matricial approach; lets say, a generalized system of multiparameter survey estimation. Practical applications of the matricial approach may result in a significant gain in computational efficiency and a better comprehension of the theory of multivariate estimation in finite populations. It is supposed that survey methodologists know the structural behavior of the population and they are able to choose a design that performs well with respect to overall characteristics of interest such as Holmberg suggests.

The structure of the paper is as follows: The second section provides some basic definitions and introduces the foundations of multiparameter estimation by means of the Horvitz-Thompson estimator. The third section is related to the estimation under the most common element sampling designs. The fourth section is related to the use of auxiliary estimation. The last section deals with a numerical example of the design and estimation applied in realistic multipurpose survey context. In that section, with the help of the R[1] `sampling` package (Tillé & Matei 2008), it is shown in detail how to produce estimations for a multiparameter survey avoiding computational loops that makes the estimation stage more difficult and slow. Proofs of the results are not shown by simplicity of the lecture. However, the reader familiarized with matrix algebra and the basic principles of survey sampling will find no problem with this.

This paper gives a comprehensive approach of joint estimation in survey sampling. Although the results of this paper are simple, it offers a powerful way of estimation in multipurpose surveys.

2. Estimating Several Parameters

Let $U = \{1, \ldots, k, \ldots, N\}$ be a finite population of N elements[2]. A probability sample s is drawn from U according to a sampling design $p(\cdot)$ that yields the first order inclusion probability of the unit k, π_k, defined as

$$\pi_k = Pr(k \in s) = \sum_{s \ni k} p(s) \tag{1}$$

[1] R is a statistical software that is very efficient in matricial computation. (R Development Core Team 2008)

[2] The population size is often not known

and the second order inclusion probability of the units k and l, defined as

$$\pi_{kl} = Pr(k, l \in s) = \sum_{s \ni k,l} p(s) \tag{2}$$

Suppose that the survey involves the study of Q characteristics of interest. Associated with the kth unit ($k \in U$) there is a vector of Q characteristics of interest, $\mathbf{y}_k = (y_{k1}, \ldots, y_{kQ})$ whose values are unknown for the entire population. In this way, the following matrix

$$\mathbf{Y}_U = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1Q} \\ \vdots & \vdots & \ddots & \vdots \\ y_{k1} & y_{k2} & \cdots & y_{kQ} \\ \vdots & \vdots & \ddots & \vdots \\ y_{N1} & y_{N2} & \cdots & y_{NQ} \end{bmatrix} = \begin{bmatrix} \mathbf{y}^1 & \mathbf{y}^2 & \ldots & \mathbf{y}^Q \end{bmatrix} \tag{3}$$

will be called the matrix of interest. Note that the element y_{kq} refers to the value of the qth characteristic of interest in the kth element, with $k \in U$ and $q = 1, \ldots, Q$. In the design based context, $\mathbf{y}^q$ is not considered a random vector, because its components are considered fixed but unknown. Hence, the values of each characteristic of interest are not necessarily continuous such as income, expenditure or weight, but also could be indicators of the membership of a population subgroup such as domains, strata or post-strata. Then, $\mathbf{Y}_U$ could be a *mixed-valued* matrix.

The objective is to estimate the Q components of the vector of totals defined by

$$\mathbf{t} = (t_1, t_2, \ldots, t_Q)' = \mathbf{Y}_U' \mathbf{1}_N \tag{4}$$

where $\mathbf{1}_N = (1, 1, \ldots, 1)'_{N \times 1}$ and $t_q = \sum_{k \in U} y_{kq}$. When the sample of size n is drawn, y_{kq} is observed ($k \in s$)and it is possible to define the following matrix

$$\mathbf{Y}_s = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1Q} \\ \vdots & \vdots & \ddots & \vdots \\ y_{k1} & y_{k2} & \cdots & y_{kQ} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nQ} \end{bmatrix} \tag{5}$$

When $s = U$, $\mathbf{Y}_U = \mathbf{Y}_s$. In this way, the matrix of inclusion probabilities in the sample is defined by

$$\Pi = \text{diag}(\pi_1, \pi_2, \ldots, \pi_n) \tag{6}$$

and the Horvitz-Thompson estimator of $\mathbf{t}$ is defined as

$$\widehat{\mathbf{t}}_\pi = \left(\widehat{t}_{1\pi}, \widehat{t}_{2\pi}, \ldots, \widehat{t}_{Q\pi}\right)' = \mathbf{Y}_s' \Pi^{-1} \mathbf{1}_n \tag{7}$$

with $\mathbf{1}_N = (1, 1, \ldots, 1)'_{n\times 1}$ and $\widehat{t}_{q\pi} = \sum_{k\in s} y_{kq}/\pi_k$ being the Horvitz-Thompson estimator of t_q. It is easy to show that $\widehat{\mathbf{t}}_\pi$ is an unbiased estimator of $\mathbf{t}$, and its covariance matrix is given by

$$\mathbf{V}\left(\widehat{\mathbf{t}}_\pi\right) = E\left(\widehat{\mathbf{t}}_\pi - \mathbf{t}\right)\left(\widehat{\mathbf{t}}_\pi - \mathbf{t}\right)' \tag{8}$$

Note that, if $N \geq q$, $\mathbf{V}\left(\widehat{\mathbf{t}}_\pi\right)$ is a positive defined symmetric matrix whose qq' element is

$$\sum_{k\in U}\sum_{l\in U} \Delta_{kl}\frac{y_{kq}}{\pi_k}\frac{y_{lq'}}{\pi_l} \tag{9}$$

with $\Delta_{kl} = \pi_{kl} - \pi_k\pi_l$ and if $s \neq U$, then it is impossible to calculate the value of the previous expression. However, if $n \geq q$, it could be estimated unbiasedly by a positive definite matrix $\widehat{\mathbf{V}}\left(\widehat{\mathbf{t}}_\pi\right)$ whose qq' element is

$$\sum_{k\in s}\sum_{l\in s} \frac{\Delta_{kl}}{\pi_{kl}}\frac{y_{kq}}{\pi_k}\frac{y_{lq'}}{\pi_l} \tag{10}$$

In some cases, the requirement of the survey is the estimation of the vector of population means given by

$$\overline{\mathbf{y}} = \frac{1}{N}\mathbf{t} \tag{11}$$

Then, an unbiased estimator of $\bar{\mathbf{y}}$ is

$$\overline{\mathbf{y}}_\pi = \frac{1}{N}\widehat{\mathbf{t}}_\pi \tag{12}$$

whose covariance matrix is estimated unbiasedly by $\frac{1}{N^2}\widehat{\mathbf{V}}\left(\widehat{\mathbf{t}}_\pi\right)$. If the population size is unknown, it can be estimated unbiasedly by using the principles of the Horvitz-Thompson estimator, such that

$$\widehat{N}_\pi = \mathbf{1}'_n\Pi^{-1}\mathbf{1}_n \tag{13}$$

Note that computational efficiency could be higher with the incorporation of the matrix approach because the estimation of several parameters is obtained by performing a matrix algebraic operation.

3. Some Sampling Designs

In this section, some examples of the estimation of several parameters, under the most used sampling designs, are given.

Example 1. Under Bernoulli sampling design, $\mathbf{t}$ is estimated unbiasedly by

$$\widehat{\mathbf{t}}_\pi = \frac{1}{\pi}\mathbf{Y}'_s\mathbf{1}_n \tag{14}$$

and its covariance matrix is estimated unbiasedly by

$$\widehat{\mathbf{V}}\left(\widehat{\mathbf{t}}_{\pi}\right) = \frac{1}{\pi}\left(\frac{1}{\pi} - 1\right)\mathbf{Y}_s\mathbf{Y}_s' \tag{15}$$

Example 2. Simple random sampling without replacement (SI) design is not the most utilized of the sampling designs, but it is useful in the last stages of a complex survey. Under this design $\mathbf{t}$ is estimated unbiasedly by

$$\widehat{\mathbf{t}}_{\pi} = \frac{N}{n}\mathbf{Y}_s'\mathbf{1}_n \tag{16}$$

And its covariance matrix is estimated unbiasedly by

$$\widehat{\mathbf{V}}\left(\widehat{\mathbf{t}}_{\pi}\right) = \frac{N^2}{n}\left(1 - \frac{n}{N}\right)\mathbf{S}_y \tag{17}$$

with $\mathbf{S}_y$, being the covariance matrix of the characteristics of interest in the sample. On the other hand, $\overline{\mathbf{y}}$ is estimated unbiasedly by

$$\overline{\mathbf{y}}_{\pi} = \frac{1}{N}\widehat{\mathbf{t}}_{\pi} = \frac{1}{n}\mathbf{Y}_s'\mathbf{1}_n \tag{18}$$

Its covariance matrix is estimated unbiasedly by the following expression

$$\widehat{\mathbf{V}}\left(\overline{\mathbf{y}}_{\pi}\right) = \frac{1}{N^2}\widehat{\mathbf{V}}\left(\widehat{\mathbf{t}}_{\pi}\right) \tag{19}$$

3.1. Estimation for Domains

If the requirements of the survey are related to the estimation of the size of a domain or the total of the characteristics of interest in such domain, the next methodological construction is required.

Let us suppose that the population is divided in D domains, such that $U = U_1, \ldots, U_d, \ldots, U_D$. Then, the indicator matrix of domain membership is defined by

$$\mathbf{Z} = \begin{bmatrix} z_{11} & \cdots & z_{1d} & \cdots & z_{1D} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ z_{k1} & \cdots & z_{kd} & \cdots & z_{kD} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ z_{n1} & \cdots & z_{nd} & \cdots & z_{nD} \end{bmatrix} \tag{20}$$

where the element

$$z_{kd} = \begin{cases} 1 & \text{if } k \in U_d, \text{ and} \\ 0 & \text{otherwise} \end{cases} \tag{21}$$

The vector of absolute domain sizes is given by

$$\mathbf{N}_d = (N_1, N_2, \ldots, N_D)' \tag{22}$$

where

$$N_d = \sum_{k \in U} z_{kd} \tag{23}$$

$\mathbf{N}_d$ is estimated by means of the Horvitz-Thompson estimator as follows

$$\widehat{\mathbf{N}}_d = \left(\widehat{N}_1, \widehat{N}_2, \ldots, \widehat{N}_D\right)' = \mathbf{Z}'\Pi^{-1}\mathbf{1}_n \tag{24}$$

and its variance is estimated unbiasedly by $\widehat{\mathbf{V}}\left(\widehat{\mathbf{N}}_d\right)$, defined similarly as in (8)

Totals of the characteristics of interest overall domains are commonly required parameters in real applications. The total of the qth variable overall D domains is given by

$$\mathbf{t}_{dq} = (t_{1q}, t_{2q}, \ldots, t_{Dq})'$$

and could be estimated by the following expression

$$\widehat{\mathbf{t}}_{dq\pi} = \left(\widehat{t}_{1q\pi}, \widehat{t}_{2q\pi}, \ldots, \widehat{t}_{Dq\pi}\right)' = (\mathbf{y}^q \mathbf{1}_D \odot \mathbf{Z})'\Pi^{-1}\mathbf{1}_n \tag{25}$$

Where $\mathbf{y}^q$ denotes the qth column of the matrix $\mathbf{Y}_s$, $\mathbf{1}_D = (1, \ldots, 1)'_{D\times 1}$ and $\odot$ denotes the Hadamard matrix product.

Example 3. Under SI design, the Horvitz-Thompson estimator for the absolute domain sizes and the total of the qth characteristic of study overall D domains are, respectively, given by

$$\widehat{\mathbf{N}}_d = (N/n)\mathbf{Z}'\mathbf{1}_n \tag{26}$$

$$\widehat{\mathbf{t}}_{dq\pi} = (N/n)(\mathbf{y}^q \mathbf{1}_D \odot \mathbf{Z})\mathbf{1}_n \tag{27}$$

3.2. Stratified Estimation

For stratified sampling designs, the finite population U is divided into H mutually exclusive strata: denoted by $U_1, \ldots, U_h \ldots, U_H$. Note that, before data collection, the membership of each element is known over the strata. This way, in each stratum a random sample is drawn. Although, it is straightforward to produce estimates by using the principles of the Horvitz-Thompson estimator, it is required to sort the population such that the matrix $\mathbf{Y}_s$ is partitioned into H blocks, as follows

$$\mathbf{Y}_s = \begin{bmatrix} \mathbf{Y}_1 \\ \vdots \\ \mathbf{Y}_h \\ \vdots \\ \mathbf{Y}_H \end{bmatrix} \tag{28}$$

where $\mathbf{Y}_h$ is a matrix that contains the values of each characteristic of interest for the elements that belong to the hth stratum, with $h = 1, \ldots, H$. Note that

$\mathbf{Y}_s \in \Re^{Hn\times Q}$ and $\mathbf{Y}_h \in \Re^{n_h\times Q}$, where given $\mathbf{n} = (n_1, \ldots, n_H)'$, then $n = \mathbf{n}'\mathbf{1}_H = n_1 + \cdots + n_H$.

As usual, the objective is to estimate the Q components of the vector of totals in the hth stratum given by

$$\mathbf{t}_h = \left(t_{1h}, t_{2h}, \ldots, t_{Qh}\right)' = \mathbf{Y}_h'\mathbf{1}_{N_h} \tag{29}$$

where N_h is the size of the hth stratum. The population total can be written as

$$\mathbf{t} = \left(t_1, t_2, \ldots, t_Q\right)' = \sum_{h=1}^{H} \mathbf{t}_h \tag{30}$$

where $\mathbf{t}_h$ is estimated unbiasedly by the following expression

$$\widehat{\mathbf{t}}_{h\pi} = \left(\widehat{t}_{1h\pi}, \widehat{t}_{2h\pi}, \ldots, \widehat{t}_{Qh\pi}\right)' = \mathbf{Y}_h'\mathbf{1}_{n_h} \tag{31}$$

where n_h is the size of the sample in the hth stratum. Of course, independence over sampling design is assumed at each stratum. This way, the estimator of the population total is given by

$$\widehat{\mathbf{t}}_{\pi} = \left(\widehat{t}_{1\pi}, \widehat{t}_{2\pi}, \ldots, \widehat{t}_{Q\pi}\right)' = \sum_{h=1}^{H} \widehat{\mathbf{t}}_h \tag{32}$$

and its variance can be written as

$$\mathbf{V}_{ST}\left(\widehat{\mathbf{t}}_{\pi}\right) = \sum_{h=1}^{H} \mathbf{V}_h\left(\widehat{\mathbf{t}}_{\pi}\right) \tag{33}$$

which is estimated unbiasedly by

$$\widehat{\mathbf{V}}_{ST}\left(\widehat{\mathbf{t}}_{\pi}\right) = \sum_{h=1}^{H} \widehat{\mathbf{V}}_h\left(\widehat{\mathbf{t}}_{\pi}\right) \tag{34}$$

Example 4. Under the stratified sampling with SI within each strata (STSI) design, the π estimator of the population total is

$$\widehat{\mathbf{t}}_{\pi} = \sum_{h=1}^{H} \frac{N_h}{n_h}\mathbf{Y}_h'\mathbf{1}_{n_h} \tag{35}$$

and its covariance matrix is estimated unbiasedly by

$$\widehat{\mathbf{V}}_{STSI}\left(\widehat{\mathbf{t}}_{\pi}\right) = \sum_{h=1}^{H} \frac{N_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right)\mathbf{S}_{yh} \tag{36}$$

with $\mathbf{S}_{yh}$, the sample covariance matrix of the study variables in the hth stratum.

4. The use of Auxiliary Information

Let us suppose that associated with the kth unit ($k \in U$) there is a vector of P auxiliary variables, $\mathbf{x}_k$. The values of the vector $\mathbf{x}_k = (x_{k1}, \ldots, x_{kP})$ are known for the entire population. Hence, the following matrix

$$\mathbf{X}_U = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1P} \\ \vdots & \vdots & \ddots & \vdots \\ x_{k1} & x_{k2} & \ldots & x_{kP} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \ldots & x_{NP} \end{bmatrix} = \begin{bmatrix} \mathbf{x}^1 & \mathbf{x}^2 & \ldots & \mathbf{x}^P \end{bmatrix} \tag{37}$$

will be called the matrix of auxiliary information.

4.1. Some Remarks

It is possible to assume that an explicit linear relationship exists between each of the characteristics of interest and the auxiliary information given by a superpopulation model ξ_q, $q = 1, \ldots, Q$, such that

$$\underset{(N\times 1)}{\mathbf{Y}^q} = \underset{(N\times P)}{\mathbf{X}}\underset{(P\times 1)}{\boldsymbol{\beta}^q} + \underset{(N\times 1)}{\boldsymbol{\varepsilon}^q}$$

The model ξ_q has the following features:

$$\begin{aligned} E_{\xi_q}(\boldsymbol{\varepsilon}^q) &= \mathbf{0} \\ V_{\xi_q}(\boldsymbol{\varepsilon}^q) &= \Sigma_q \end{aligned} \tag{38}$$

Σ_q establishes the variance structure of the vector $\boldsymbol{\varepsilon}^q$. The previous relationships can also be described by a joint superpopulation model ξ such that

$$\underset{(N\times Q)}{\mathbf{Y}} = \underset{(N\times P)}{\mathbf{X}}\underset{(P\times Q)}{\boldsymbol{\beta}} + \underset{(N\times Q)}{\boldsymbol{\varepsilon}}$$

Note that this approach suggests that $\mathbf{Y}$, $\mathbf{X}$ and $\boldsymbol{\varepsilon}$ are random matrices (Gupta & Nagar 2000) defined in the superpopulation model ξ, of which $\mathbf{Y}_U$ and $\mathbf{X}_U$ are supposed to be the outcomes. More precisely, the model ξ has the following features:

$$\begin{aligned} E_\xi(\boldsymbol{\varepsilon}) &= \underset{(N\times Q)}{\mathbf{0}} \\ V_\xi(\operatorname{vec}\boldsymbol{\varepsilon}) &= \underset{(NQ\times NQ)}{\boldsymbol{\Sigma}} = \operatorname{diag}\big(\Sigma_1, \Sigma_2, \ldots, \Sigma_Q\big) \end{aligned} \tag{39}$$

Note that the subindex ξ refers to the expectation under the assumed working model. In practical situations, it is often common to take $\Sigma_q = \sigma_q^2 \operatorname{diag}(c_{1q}, \ldots, c_{Nq})$, where $c_{kq} = f_q(x_{k1}, \ldots, x_{kP})$ and f_q is a real-valued function.

The problem of estimating $\boldsymbol{\beta}$ is considered briefly. Let $D(\mathbf{X})$ be a translation-invariant measure of dispersion such that $D(\mathbf{X}+\mathbf{K}) = D(\mathbf{X})$, with $\mathbf{K}$ a matrix of constants. Then the estimation of $\mathbf{B}$ will be the matrix that minimizes the measure of dispersion. Particularly, $D(\cdot)$ could be given by the multivariate total variance defined as

$$trace(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}) \tag{40}$$

By using the least squares method, (40) is minimized by the following choice

$$\mathbf{B} = \left(\mathbf{B}_1, \mathbf{B}_2, \ldots, \mathbf{B}_Q\right) \tag{41}$$

where

$$\mathbf{B}_q = \left(\mathbf{X}_U'\Sigma_q^{-1}\mathbf{X}_U\right)^{-1}\left(\mathbf{X}_U'\Sigma_q^{-1}\mathbf{Y}_U\right) \tag{42}$$

4.2. Classical Information

In real applications, just one sample is drawn from a finite population and it is not possible to compute $\mathbf{B}$. Then, it must be estimated with the information available in the sample. It can be shown that an asymptotically unbiased estimator of $\mathbf{B}$ is given by

$$\widehat{\mathbf{B}} = \left(\widehat{\mathbf{B}_1}, \widehat{\mathbf{B}_2}, \ldots, \widehat{\mathbf{B}_Q}\right) \tag{43}$$

where

$$\widehat{\mathbf{B}_q} = (\mathbf{X}_s'\mathbf{A}_q^{-1}\mathbf{X}_s)^{-1}(\mathbf{X}_s'\mathbf{A}_q^{-1}\mathbf{Y}_s) \tag{44}$$

$q = 1, \ldots, Q$, $\mathbf{X_s}$ is defined similarly to (5) and

$$\mathbf{A}_q = \Pi^{1/2}\Sigma_q\Pi^{1/2} \tag{45}$$

Hence, the multiple generalized regression estimator of the population total is defined by

$$\widehat{\mathbf{t}}_{Mgreg} = \widehat{\mathbf{t}}_{\mathbf{y}\pi} + \widehat{\mathbf{B}}'\left(\mathbf{t}_{\mathbf{x}} - \widehat{\mathbf{t}}_{\mathbf{x}\pi}\right) \tag{46}$$

with, $\widehat{\mathbf{t}}_{\mathbf{y}\pi}$, $\widehat{\mathbf{t}}_{\mathbf{x}\pi}$ the Horvitz-Thompson estimators of $\mathbf{t_y}$ and $\mathbf{t_x}$, respectively. Note that $\widehat{\mathbf{B}}_q$ could also be written as

$$\begin{aligned}\widehat{\mathbf{B}}_q &= (\mathbf{X}_s'\mathbf{D}_\lambda\mathbf{X}_s)^{-1}\mathbf{X}_s\mathbf{D}_\lambda\mathbf{Y}_s \\ &= \left(\sum_{k\in s}\mathbf{x}_k\lambda_k^q\mathbf{x}_k'\right)^{-1}\left(\sum_{k\in s}\mathbf{x}_k\lambda_k^q\mathbf{y}_k'\right)\end{aligned} \tag{47}$$

where $\mathbf{D}_\lambda = \text{diag}(\lambda_1^q, \ldots, \lambda_n^q)$ and λ_k^q some real-valued function of the probabilities of inclusion and the auxiliary information. Note that the model ξ serves as a vehicle for finding an appropriate general regression estimator. Once the estimator is found, the model is no longer of use. The properties of the multiple generalized regression estimator (expectation and variance) are still derived from a design based perspective.

4.2.1. Some Particular Cases

All of the following cases are enounced under a general assumption. Then, several special cases of the multiple generalized regression estimator are distinguished, depending on the choice of the values of λ_k.

- If $P = 1$, $\mathbf{x}_k = x_k$, and $\lambda_k^q = (\pi_k x_k)^{-1}$, we have the ratio estimator for each characteristic of interest,
- If $P = 2$, $\mathbf{x}_k = (1, x_k)'$, and $\lambda_k^q = (\pi_k)^{-1}$, we have the classical simple regression estimator,
- If $P = M$(number of post-strata), $\mathbf{x}_k = \delta_k = (0, \ldots 0, 1, 0, \ldots, 0)'$, and $\lambda_k^q = (\pi_k)^{-1}$, where δ_k represents M dummy variables (each dummy represents a post-stratum membership), we have the post-stratified estimator.

The multiple generalized regression estimator can be written in a simplified form given by

$$\widehat{\mathbf{t}}_{Mgreg} = (\mathbf{W}' \odot \mathbf{Y}_s')\mathbf{1}_n \tag{48}$$

where

$$\mathbf{W} = \begin{bmatrix} w_1^1 & w_1^2 & \ldots & w_1^Q \\ \vdots & \vdots & \ddots & \vdots \\ w_k^1 & w_k^2 & \ldots & w_k^Q \\ \vdots & \vdots & \ddots & \vdots \\ w_n^1 & w_n^2 & \ldots & w_n^Q \end{bmatrix} \begin{bmatrix} \mathbf{w}^1 & \mathbf{w}^2 & \ldots & \mathbf{w}^Q \end{bmatrix} \tag{49}$$

Note that $\mathbf{w}^q = \left(w_1^q, \ldots, w_k^q, \ldots, w_n^q\right)'$ is a vector of weights such that

$$w_k^q = \frac{1}{\pi_k}\left(1 + \lambda_k^q \mathbf{x}_k' \left(\sum_{k \in s} \mathbf{x}_k \lambda_k^q \mathbf{x}_k'\right)^{-1} (\mathbf{t}_\mathbf{x} - \widehat{\mathbf{t}}_{\mathbf{x}\pi})\right) \tag{50}$$

These weights are often called the calibration weights and they reproduce the vector of totals $\mathbf{t}_\mathbf{x}$ when they are applied to the auxiliary information. Then, $\mathbf{W}$ is called the calibration matrix. It is not difficult to show that

$$\sum_{k \in s} w_k^q \mathbf{x}_k = \mathbf{X}_s' \mathbf{w}^q = \mathbf{t}_\mathbf{x} \tag{51}$$

for all $q = 1, \ldots, Q$. It is very interesting to see that $\mathbf{t}_\mathbf{x}$ is calibrated under different choices of w^q's. On the other hand, note that

$$\mathbf{w}^q = \Pi^{-1}\mathbf{1}_n + \mathbf{A}_q \mathbf{X}_s \left(\mathbf{X}_s' \mathbf{A}_q \mathbf{X}_s\right)^{-1} \left(\mathbf{t}_\mathbf{x} - \widehat{\mathbf{t}}_{\mathbf{x}\pi}\right) \tag{52}$$

In the post-stratified estimation, a generalized inverse must be used; however, the multiple generalized regression estimator is invariant to the choice of the inverse.

4.3. Joint Auxiliary Information

The least squares method is not the only way to obtain a multiple regression estimator. In this subsection, it is supposed that a joint information matrix could be constructed by means of

$$\mathbf{V} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1Q} & x_{11} & x_{12} & \cdots & x_{1P} \\ y_{21} & y_{22} & \cdots & y_{2Q} & x_{21} & x_{22} & \cdots & x_{2P} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nQ} & x_{n1} & x_{n2} & \cdots & x_{nP} \end{bmatrix} \tag{53}$$

The estimator of the vector of totals of both, characteristics of interest and auxiliary information, is given by $\widehat{\mathbf{t}}_{\mathbf{v}\pi}$, defined as

$$\widehat{\mathbf{t}}_{\mathbf{v}\pi} = \mathbf{V}'\Pi^{-1}\mathbf{1}_n \tag{54}$$

Then, let us suppose that $\widehat{\mathbf{t}}_{\mathbf{v}\pi}$ has a multivariate normal distribution with mean

$$E\left(\widehat{\mathbf{t}}_{\mathbf{v}\pi}\right) = \mathbf{t}_{\mathbf{v}} = \left(\mathbf{t}'_{\mathbf{Y}\pi}, \mathbf{t}'_{\mathbf{X}\pi}\right)'$$

and covariance matrix given by

$$V\left(\widehat{\mathbf{t}}_{\mathbf{v}\pi}\right) = \begin{bmatrix} V\left(\widehat{\mathbf{t}}_{\mathbf{y}\pi}\right) & C\left(\widehat{\mathbf{t}}_{\mathbf{y}\pi}, \widehat{\mathbf{t}}_{\mathbf{x}\pi}\right) \\ C\left(\widehat{\mathbf{t}}_{\mathbf{y}\pi}, \widehat{\mathbf{t}}_{\mathbf{x}\pi}\right) & V\left(\widehat{\mathbf{t}}_{\mathbf{x}\pi}\right) \end{bmatrix}$$

where $V\left(\widehat{\mathbf{t}}_{\mathbf{y}\pi}\right)$ is a symmetric matrix such that the jth element of its diagonal is given by the variance of $\widehat{t}_{y_j\pi}$

$$V\left(\widehat{t}_{y_j\pi}\right) = \sum\sum_{U} \Delta_{kl} \frac{y_{jk}}{\pi_k} \frac{y_{jl}}{\pi_l}$$

and the off-diagonal element ij is given by the covariance of $\widehat{t}_{y_i\pi}$ and $\widehat{t}_{y_j\pi}$,

$$C\left(\widehat{t}_{y_i\pi}, \widehat{t}_{y_j\pi}\right) = \sum\sum_{U} \Delta_{kl} \frac{y_{ik}}{\pi_k} \frac{y_{jl}}{\pi_l}$$

$V\left(\widehat{\mathbf{t}}_{\mathbf{x}\pi}\right)$ is similarly defined, and $C\left(\widehat{\mathbf{t}}_{\mathbf{y}\pi}, \widehat{\mathbf{t}}_{\mathbf{x}\pi}\right)$, not necessarily symmetric, is such that the element ij is given by the covariance of $\widehat{t}_{y_i\pi}$ and $\widehat{t}_{x_j\pi}$

$$C\left(\widehat{t}_{y_i}\pi, \widehat{t}_{x_j}\pi\right) = \sum\sum_{U} \Delta_{kl} \frac{y_{ik}}{\pi_k} \frac{x_{jl}}{\pi_l}$$

According to the multivariate inference for normal populations, the conditional distribution of $\widehat{\mathbf{t}}_{\mathbf{Y}\pi}$ given $\widehat{\mathbf{t}}_{\mathbf{X}\pi}$ follows a multivariate normal distribution with a conditional mean given by

$$E\left(\widehat{\mathbf{t}}_{\mathbf{y}\pi} \mid \widehat{\mathbf{t}}_{\mathbf{x}\pi}\right) = \mathbf{t}_{\mathbf{y}\pi} + C\left(\widehat{\mathbf{t}}_{\mathbf{y}\pi}, \widehat{\mathbf{t}}_{\mathbf{x}\pi}\right)\left(V\left(\widehat{\mathbf{t}}_{\mathbf{x}\pi}\right)\right)^{-1}\left(\mathbf{t}_{\mathbf{x}} - \widehat{\mathbf{t}}_{\mathbf{x}\pi}\right) \tag{55}$$

and a conditional variance given by

$$V\left(\widehat{\mathbf{t}}_{\mathbf{y}\pi} \mid \widehat{\mathbf{t}}_{\mathbf{x}\pi}\right) = V\left(\widehat{\mathbf{t}}_{\mathbf{y}\pi}\right) - C\left(\widehat{\mathbf{t}}_{\mathbf{y}\pi}, \widehat{\mathbf{t}}_{\mathbf{x}\pi}\right)\left(V\left(\widehat{\mathbf{t}}_{\mathbf{x}\pi}\right)\right)^{-1} C\left(\widehat{\mathbf{t}}_{\mathbf{x}\pi}, \widehat{\mathbf{t}}_{\mathbf{y}\pi}\right) \tag{56}$$

Note that 55 and (56) are unbiasedly estimated by

$$\begin{aligned} \widehat{\mathbf{t}}_{\mathbf{y}} &= \widehat{\mathbf{t}}_{\mathbf{y}\pi} + \widehat{C}\left(\widehat{\mathbf{t}}_{\mathbf{y}\pi}, \widehat{\mathbf{t}}_{\mathbf{x}\pi}\right)\left(\widehat{V}\left(\widehat{\mathbf{t}}_{\mathbf{x}\pi}\right)\right)^{-1}\left(\mathbf{t}_{\mathbf{x}} - \widehat{\mathbf{t}}_{\mathbf{x}\pi}\right) \\ &= \widehat{\mathbf{t}}_{\mathbf{y}\pi} + \widehat{\mathbf{B}}\left(\mathbf{t}_{\mathbf{x}} - \widehat{\mathbf{t}}_{\mathbf{x}\pi}\right) \end{aligned} \tag{57}$$

and,

$$\widehat{V}\left(\widehat{\mathbf{t}}_{\mathbf{y}}\right) = V\left(\widehat{\mathbf{t}}_{\mathbf{Y}\pi}\right) - C\left(\widehat{\mathbf{t}}_{\mathbf{Y}\pi}, \widehat{\mathbf{t}}_{\mathbf{X}\pi}\right)\left(V\left(\widehat{\mathbf{t}}_{\mathbf{X}\pi}\right)\right)^{-1} C\left(\widehat{\mathbf{t}}_{\mathbf{X}\pi}, \widehat{\mathbf{t}}_{\mathbf{Y}\pi}\right) \tag{58}$$

respectively. Note that (58) looks like the multiple regression estimator. However, its slope, $\widehat{\mathbf{B}}$, is different. While the slope of the latter is given by the use of the least squares method, the slope of the first corresponds to a set of multiple regressions of $\mathbf{X}$ over $\mathbf{Y}$. This estimator of the total should be called a multiple optimal regression estimator and has been studied by Cassady & Valiant (1993) in a model based context for the estimation of a total of a single characteristic of interest.

5. A Numerical Example

In this section, an example of the multiparameter approach is considered. In the design stage, an optimal sampling design must be chosen and the Holmberg's approach (Holmberg 2002*b*) is used. In the estimation stage the matricial approach, proposed in the preceding sections, is used. Both, the design and estimation stages are implemented using the statistical software R. Specifically the package `sampling` is used in selecting the sample and estimating several parameters in domains of interest.

For this purpose, a realistic population is used (the population of Swedish municipalities MU281 available in appendix B in Särndal et al. (1992)). This way, it is possible to plan a multiparameter survey, where auxiliary variables and domains are supplied, and where it is feasible to have some kind of beliefs about the validity of the relationships between the study variables and the auxiliary variables. The main issue of this section is not focused on planning a perfect sampling design, but on illustrating the performance (design and estimation) of a multiparameter survey.

The study variables are:

y_1	=	P85	(1985 population)
y_2	=	RMT85	(Revenues from the 1985 municipal taxation)
y_3	=	REV84	(Real state values according to 1984 assessment)

The auxiliary variables are:

x_1 = P75 (1975 population)
x_2 = S82 (Total number of seats in the municipal council in 1982)

For domain estimation the following variable is used: z=REG (Geographic region indicator)

The following code could be used to specify the previous features of the survey.

```
> library(sampling)
> data(MU284)
> MU281 <- MU284[MU284$RMT85$<=3000,]
> attach(MU281)

> Y1 <- P85
> Y2 <- RMT85
> Y3 <- REV84
> X1 <- P75
> X2 <- S82
> Z  <- REG
```

In order to have some kind of idea about the resulting estimations, it is useful to know the totals of the study variables and the auxiliary variables. So, the following code could be used to obtain such totals.

```
> Ty <- c(sum(Y1),sum(Y2),sum(Y3))
> Tx <- c(N,sum(X1),sum(X2))

> Ty
[1]   7033  53151 757246
> Tx
[1]   281  6818 13257
```

5.1. Design Stage under Holmberg's Approach

Let us suppose that the importance of the three study variables is the same. A brief summary of the Holmberg's approach is as follows:

1. For each of the study variables, the survey statistician must propose an optimal sampling design, $p_q(\cdot)$, such that the expected sample size is $E_p(n_s) = n_q$. Of course, note that all of the Q designs may differ; but even more, the sample sizes may not be necessarily the same. In the traditional way, the statistician should choose a compromise design that works well for all parameters to be estimated.

   In the MU281 example, the population size is $N = 281$, and let us suppose that the statistician assumes to take a sample size of 100 units for all of the three sampling designs. This specifications in R are as simple as follows

   ```
   > N <- 281; n <- 100
   ```

2. The design $P_q(\cdot)$ induces a vector of size N of inclusion probabilities for every unit in the population. The inclusion probabilities must take the following form (Holmberg 2002*b*, eq. 6)

$$\pi_{qk} = n_q \frac{\sigma_{qk}}{\sum_{k\in s} \sigma_{qk}}$$

with σ_{qk} size measures (usually, but not necessarily, linked to a regression model). Optimality is obtained if $\pi_{qk} \propto \sigma_{qk}$. Note that if the optimal sampling design for the qth study variable is SI, then $\sigma_{qk} = 1$ for all $k \in U$. On the other hand, by choosing $\sigma^2_{qk} = \sigma^2_q x^{\gamma_q}_{qk}$, where σ^2_q is a constant and x_{qk} is the value of the kth unit for some auxiliary variable, then the optimal sampling design is a probability proportional to size (πPS) with $\pi_{qk} \propto x^{\gamma_q/2}_{qk}$.

Let us suppose that the statistician is confident that the best optimal sampling designs, for the study variables of the MU281 population, are: For y_1, a πPS design with $\pi_{1k} \propto x^{0.7}_{1k}$. For y_2, a πps design with $\pi_{2k} \propto x_{1k}$. For y_3, a SI design with $\pi_{3k} = 100/281$. The R code is as follows

```
> sigy1 <- sqrt(X1^(1.4))
> sigy2 <- sqrt(X1^(2))
> sigy3 <- rep(1,N)

> pik1 <- n*sigy1/(sum(sigy1))
> pik2 <- n*sigy2/(sum(sigy2))
> pik3 <- n*sigy3/(sum(sigy3))
```

3. Based on the criterion of minimum anticipated overall relative efficiency loss (ANOREL), the optimal sample size for the multiparameter case is given by

$$n^* \geq \frac{\left(\sum_{k\in U} \sqrt{a_{qk}}\right)^2}{(1+c)Q + \sum_{k\in U} a_{qk}}$$

where

$$a_{qk} = \sum_{q=1}^{Q} \frac{\sigma^2_{qk}}{\sum_{k\in U} \left(\frac{1}{\pi_{qk}} - 1\right) \sigma^2_{qk}}$$

and c is the maximum error allowed based on the ANOREL criterion. Note that in practice, σ^2_{qk} is unattainable and it must be written as a function of an auxiliary variable. Holmberg (2002*b*) claims that subject knowledge, guesses, or previous survey estimates can be used as planning values for this quantity.

In the MU281 example, the optimal sample size based on the ANOREL criterion under the multiparameter case is $n^* = 108$. The R code is as follows

```
> a1 <-  sigy1^2/(sum(((1/pik1)-1)*sigy1^2))
> a2 <-  sigy2^2/(sum(((1/pik2)-1)*sigy2^2))
> a3 <-  sigy3^2/(sum(((1/pik3)-1)*sigy3^2))
> aqk <- a1+a2+a3

> n.st <- ((sum(sqrt(aqk)))^2)/((1+0.03)*3+(sum(aqk)))
> n.st <- as.integer(n.st)
[1] 108
```

4. Once the sample size is computed, a new vector of optimal inclusion probabilities for all of the study variables is created. This vector is induced by a general sampling design, which minimizes the anticipated overall relative efficiency loss, and it is given by

$$\pi_{(opt)k} = \frac{n^* \sqrt{a_{qk}}}{\sum_{k \in U} \sqrt{a_{qk}}} \tag{59}$$

   In the MU281 example, the optimal vector of inclusion probabilities is given by the following code

```
> pikopt <- n.st*sqrt(aqk)/sum(sqrt(aqk))
> sum(pikopt) == n.st
[1] TRUE
```

5. In most cases, the resulting $\pi_{(opt)k}$ is a vector of unequal inclusion probabilities. In this situation a πps selection scheme must be performed in order to select the sample.

   In the MU281 example, let us suppose that the statistician recommends an order sampling (unequal probabilities, without replacement and fixed sample size) to be performed. The function `UPopips` of the `sampling` package selects a sample with the previous features and after the sample has been drawn, the function `getdata` extracts the observed data.

```
> sam <- UPopips(pikopt,"exponential")
> getdata(MU281,sam)
    LABEL P85 P75 RMT85 CS82 SS82 S82 ME84 REV84 REG CL
       74  17  18   113    8   20  49  784  1733   3 13
      184  12  11    82    6   25  41  646   935   6 33

       ...
```

5.2. Estimation Stage under Multiparameter Approach

Once the sample is selected, the statistician is faced with the estimation of several parameters of interest. It is possible to write a code of estimation for

every single study variable (the traditional way) or it is possible to write a simple matricial code using the proposed approach in this paper. Even though it is not the issue here, the gain in computational efficiency is not negligible. Besides, if the design stage is planned in a multiparameter context, then the estimation stage should be carried out in the same way.

In the MU281 example, with the new optimal inclusion probabilities, $\pi_{(opt)k}$, the HT estimator of the total of study variables, $\mathbf{t_y}$, the total of auxiliary variables, $\mathbf{t_x}$, and population size, N, given by (7) and (13), is computed by means of the following code:

```
> Ys <- cbind(Y1,Y2,Y3)[sam,]
> Xs <- cbind(1,X1,X2)[sam,]
> PI <- diag(pikopt[sam])
> ones <- rep(1,n.st)

> TyHT <- t(Ys)%*%solve(PI)%*%ones
> TxHT <- t(Xs)%*%solve(PI)%*%ones
> NHT <- t(ones)%*%solve(PI)%*%ones
```

The result of the previous computation is a vector of estimated totals; in particular, the HT estimation of the total of study variables is given by

```
> TyHT
         [,1]
Y1   6603.514
Y2  49078.942
Y3 719565.860
```

If a domain is involved in the estimation stage, the matricial approach gives a simple and comprehensive method of estimation. In the MU281 example, the domain of interest corresponds to the `REG` variable which has 8 categories or geographic regions. Then, a discriminated estimation for all of the three study variables is needed for each domain. By using the `disjunctive` function of the `sampling` package, it is possible to create the indicator matrix of domain membership given by (20) and to obtain the corresponding estimates of (24) and (25).

```
> Z <- disjunctive(Z)[sam,]
> NdHT <- t(Z)%*%solve(PI)%*%ones
> Ty1d <- t(Ys[,1]*Z)%*%solve(PI)%*%ones
> Ty2d <- t(Ys[,2]*Z)%*%solve(PI)%*%ones
> Ty3d <- t(Ys[,3]*Z)%*%solve(PI)%*%ones
```

It is feasible to summarize the estimations by means of a simple data frame, as follows:

```
> TydHT <- data.frame(NdHT,Ty1d,Ty2d,Ty3d)
```

```
> TydHT
      NdHT      Ty1d       Ty2d        Ty3d
1    24.17   1045.85     8711.74   109849.38
2    51.50    895.17     6707.68    90645.23
3    29.86    594.90     4449.43    60348.27
4    49.13   1072.34     7235.20   104268.22
5    64.01   1394.64     9930.63   142899.69
6    40.50    730.81     5521.88    78599.09
7     6.44    207.89     1560.50    22637.41
8    51.79    661.88     4961.85   110318.56
```

If the statistician suspects that a model-assisted approach can be used, then a relationship between the study variables and the auxiliary information must be established by means of a working model. In the MU281 example, there are three models, ξ_q (q=1,2,3), involved in a general superpopulation model ξ. The relationships are as follows

$$Y_q = \beta_{q0} + \beta_{q1} X_1 + \beta_{q2} X_2 + \boldsymbol{\varepsilon}_i \qquad q = 1, 2, 3 \tag{60}$$

Note that $E_{\xi_i}(\boldsymbol{\varepsilon}_i) = 0$ and the structure of variance of the preceding models is induced by step two of the design stage and it is given by

$$\begin{aligned}
\Sigma_1 &= \sigma_1^2 \,\mathrm{diag}(x_{11}, x_{12}, \ldots, x_{1N})^{1.4} \\
\Sigma_2 &= \sigma_2^2 \,\mathrm{diag}(x_{11}, x_{12}, \ldots, x_{1N})^{2} \\
\Sigma_3 &= \sigma_3^2 \mathbf{I}_{N\times N}
\end{aligned}$$

Then, the general model takes the following form

$$\begin{bmatrix} Y_{11} & Y_{21} & Y_{31} \\ Y_{12} & Y_{22} & Y_{32} \\ \vdots & \vdots & \vdots \\ Y_{1N} & Y_{2N} & Y_{3N} \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{21} \\ 1 & X_{12} & X_{22} \\ \vdots & \vdots & \vdots \\ 1 & X_{1N} & X_{2N} \end{bmatrix} \begin{bmatrix} \beta_{10} & \beta_{20} & \beta_{30} \\ \beta_{11} & \beta_{21} & \beta_{31} \\ \beta_{12} & \beta_{22} & \beta_{32} \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} & \varepsilon_{21} & \varepsilon_{31} \\ \varepsilon_{12} & \varepsilon_{22} & \varepsilon_{32} \\ \vdots & \vdots & \vdots \\ \varepsilon_{1N} & \varepsilon_{2N} & \varepsilon_{3N} \end{bmatrix}$$

In this way, the estimation of the finite population matrix of regression coefficients, involving the variance structure of each model, given by (43) is computed by means of the following code

```
> A1 <- diag(pikopt[sam]*Xs[,2]^(1.4))
> B1 <- (solve(t(Xs)%*%A1%*%Xs))%*%(t(Xs)%*%A1%*%Ys[,1])
> A2 <- diag(pikopt[sam]*Xs[,2]^(2))
> B2 <- (solve(t(Xs)%*%A2%*%Xs))%*%(t(Xs)%*%A2%*%Ys[,2])
> A3 <- diag(pikopt[sam])
> B3 <- (solve(t(Xs)%*%A3%*%Xs))%*%(t(Xs)%*%A3%*%Ys[,3])
```

```
> B <- matrix(c(B1,B2,B3),ncol=3,nrow=3)
> B
                [,1]            [,2]          [,3]
[1,]     -1.20582067     -25.5012341     868.01938
[2,]      1.05356147       8.3134903     104.90848
[3,]      0.01756820       0.4836855     -15.78483
```

The next step is to implement the multiple generalized regression estimator given by (46). The computational code requires just a line and it is given by

```
> TyMgreg <- TyHT+t(B)%*%(Tx - TxHT )
> TyMgreg
         [,1]
Y1   7079.411
Y2  53028.236
Y3 750689.737
```

The multiple generalized regression estimator can take different forms. Among others, it could be written in the simplified form given by (48). However, it is necessary to compute the calibration matrix given by (49). In R, it can be done as follows

```
> w1 <- solve(PI)%*%ones + (A1%*%Xs)%*%(solve(t(Xs)%*%A1%*%Xs))%*%(Tx - TxHT)
> w2 <- solve(PI)%*%ones + (A2%*%Xs)%*%(solve(t(Xs)%*%A2%*%Xs))%*%(Tx - TxHT)
> w3 <- solve(PI)%*%ones + (A3%*%Xs)%*%(solve(t(Xs)%*%A3%*%Xs))%*%(Tx - TxHT)
> W <- cbind(w1,w2,w3)

> TyMgreg <- t(W*Ys)%*%ones
> TyMgreg
         [,1]
Y1   7079.411
Y2  53028.236
Y3 750689.737
```

The calibration principle given by (51) can be verified easily for each column of the calibration matrix. Particularly for the second column this principle remains.

```
> t(w2)%*%Xs
           X1    X2
[1,] 281 6818 13257
```

This section has shown how to perform a multiparameter survey by using the Holmberg's approach, in the design stage, and the matricial approach, in the estimation stage, as proposed in this paper.

6. Conclusions

When planning a survey, the traditional way focuses the sampling design on a single study variable, which is insufficient for the practising survey statistician,

having to deal with several parameters of interest. In this paper, a potentially useful approach in the estimation stage is presented by means of the matricial perspective. The author stresses that more attention should be paid to the joint estimation of the parameters of interest in multipurpose surveys. This way, significant advantages will be achieved, in practical and theoretical aspects, through a comprehensive approach of joint estimation in survey sampling that offers a powerful way of estimation in multipurpose surveys.

Apart from the computational advantages, this approach should be used in order to introduce advanced topics in survey sampling such as the generalized weight share method proposed by Lavallée & Caron (2001).

Further work should be focused on the interpretation and use of covariances among the estimators of the parameters of interest as a vehicle for improving the planning, execution and estimation in business surveys.

[Recibido: agosto de 2008 — Aceptado: febrero de 2009]

References

Cassady, R. J. & Valiant, R. (1993), 'Conditional Properties of Poststratified Estimators under Normal Theory', *Survey Methodology* **19**, 183–192.

Gupta, A. K. & Nagar, D. K. (2000), *Matrix Variate Distributions*, Chapmann and Hall, New York, United States.

Holmberg, A. (2002*a*), 'A Multiparameter Perspective on the Choice of Sampling Design in Surveys', *Statistics in Transition* **5**, 969–994.

Holmberg, A. (2002*b*), On the Choice of Sampling Design under GREG Estimation in Multiparameter Surveys, Technical Report 1, *RD Department, Statistics Sweden*, SE-701 89 Örebro, Sweden.

Lavallée, P. & Caron, P. (2001), 'Estimation Using the Generalized Weigth Share Method: The use of Record Linkage', *Survey Methodology* **27**, 155–169.

R Development Core Team (2008), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
*http://www.R-project.org

Särndal, C. E., Swensson, B. & Wretman, J. (1992), *Model Assisted Survey Sampling*, Springer, New York, United States.

Tillé, Y. & Matei, A. (2008), *Sampling: Survey Sampling*. R package version 2.0.