Published in final edited form as:

J Stat Theory Pract. 2014 October 2; 8(4): 772–791. doi:10.1080/15598608.2013.856358.

Estimation of finite population parameters with auxiliary information and response error [★]

L. M. González^{a,b}, J. M. Singer^c, and E.J. Stanek III^d

L. M. González: lgonzalezg@unal.edu.co; J. M. Singer: jmsinger@ime.usp.br; E.J. Stanek: stanek@schoolph.umass.edu

^aDepartamento de Estadística, Universidad Nacional de Colombia, Bogotá, Colombia

^bTelephone: 57-1-3165000 Ext. 13210. Fax number: 57-1-3165327. Postal address: Ciudad Universitaria, edificio 405, oficina 315, Bogotá, Colombia

^cDepartamento de Estatística, Universidade de São Paulo, São Paulo, Brazil

^dDepartment of Public Health, University of Massachusetts, Amherst, USA

Abstract

We use a finite population mixed model that accommodates response error in the survey variable of interest and auxiliary information to obtain optimal estimators of population parameters from data collected via simple random sampling. We illustrate the method with the estimation of a regression coefficient and conduct a simulation study to compare the performance of the empirical version of the proposed estimator (obtained by replacing variance components with estimates) with that of the least squares estimator usually employed in such settings. The results suggest that when the auxiliary variable distribution is skewed, the proposed estimator has a smaller mean squared error.

Keywords

BLUP; optimal estimator; random permutation model; regression coefficient

1. Introduction

Estimation or prediction of latent values or of functions of such quantities is a theme with increasing interest in many fields, with special mention to Education and Psychiatry. The ability of students in Mathematics or some psychological trait of patients with different mental disorders are typical examples. In general, we observe the sum of the latent values and response error (through the scores in exams, for example) instead of the latent values themselves. Different approaches, mostly model-based, to this problem have been proposed in the literature [see Baker (1992) and Hambleton and Cook (1997), for example]. Although they have many advantages, model misspecification has been a source of criticisms. We

[★]This work was developed with financial support from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), Brazil, Universidad Nacional de Colombia, Colombia, and National Institute of Health (NIH-PHS-R01-HD36848, R01-HL071828-02), USA. The authors wish to thank Viviana Lencina, Wenjun Li, Silvina San Martino and Bo Xu for helpful comments.

propose a simple design-based alternative under which parameters are clearly defined and response error is taken into account. Our proposal is motivated by the following example.

Each year, a population of about 1000 students are admitted as freshmen in Engineering, Mathematics or Physics careers at the University of São Paulo, Brazil according to their scores on an exam (FUVEST) designed to evaluate their performance in different disciplines. In many instances, one might be interested in estimating some function of certain latent values associated to this population on the basis of a simple random sample selected therefrom. Typical examples of such functions are the mean ability in Mathematics or a regression coefficient between the Mathematics score in the FUVEST exam and the ability in Mathematics (the latent value).

Let y_s and x_s be fixed values representing the ability in Mathematics, the survey variable of interest, and the score in the FUVEST Mathematics exam, the auxiliary variable, respectively, for student s in the finite population of N students and let

$$B = \sum_{s=1}^N \frac{(y_s - \mu_y)(x_s - \mu_x)}{\sum_{s=1}^N (x_s - \mu_x)^2}, \quad (1)$$

with $\mu_y = N^{-1} \sum_{s=1}^N y_s$ and $\mu_x = N^{-1} \sum_{s=1}^N x_s$, denote the regression coefficient between the two variables. A brief discussion on the interpretation of B is given in Appendix C.

Estimation of B based on a simple random sample of size n obtained without replacement from a finite population has been considered by several authors. In particular, Kish and Frankel (1974) and Särndal, Swensson, and Wretman (1992) attacked the problem from a design-based point of view, while Fuller (1975), Holt, Smith, and Winter (1980), Pfeiffermann and Smith (1985), Bolfarine and Zacks (1992) and Bolfarine, Zacks, Elian, and Rodrigues (1994) approached it from a superpopulation point of view. Under both approaches, the ordinary least squares estimator, $\hat{B}_{OLS}$, is the suggested solution. For simple random samples, the OLS estimator corresponds to the Horvitz-Thompson estimator (π -estimator)¹. None of these approaches, however, take response error in the survey variables of interest into account. We consider an alternative estimator of such functions of latent values derived under the finite population mixed model suggested by Stanek, Singer & Lencina (2004) with the incorporation of response error as in Stanek & Singer (2004) and auxiliary information as in Li, Stanek III, and Singer (2012).

¹If π_s is the inclusion probability of unit s in a sample and $\mathcal{S}$ is the set of realized unit labels (units in the sample), the π -estimator of B is

$$\hat{B}_\pi = \left\{ \sum_{s \in \mathcal{S}} \pi_s^{-1} (x_s - \tilde{x}_\pi)^2 \right\}^{-1} \left\{ \sum_{s \in \mathcal{S}} \pi_s^{-1} (x_s - \tilde{x}_\pi) (y_s - \tilde{y}_\pi) \right\},$$

where $\tilde{x}_\pi = N^{-1} \sum_{s \in \mathcal{S}} \pi_s^{-1} x_s$ and $\tilde{y}_\pi = N^{-1} \sum_{s \in \mathcal{S}} \pi_s^{-1} y_s$. The terms $\tilde{x}_\pi$ and $\tilde{y}_\pi$ are the π -estimators of μ_x and μ_y , respectively. Under simple random sampling without replacement, $\pi_s = n/N$ for all s and then $\hat{B}_\pi = \hat{B}_{OLS} = \left\{ \sum_{s \in \mathcal{S}} (x_s - \bar{x})^2 \right\}^{-1} \left\{ \sum_{s \in \mathcal{S}} (x_s - \bar{x})(y_s - \bar{y}) \right\}$ where $\bar{x} = n^{-1} \sum_{s \in \mathcal{S}} x_s$ and $\bar{y} = n^{-1} \sum_{s \in \mathcal{S}} y_s$ [see Särndal, Swensson, and Wretman (1992), for example].

In Section 2, we describe the finite population mixed model. In Section 3, we derive estimators of linear combinations of the latent values (of which B is a special case). In Section 4, we present numerical examples to compare the performance of the proposed estimator of B with that of the ordinary least squares estimator, B_{OLS} . We also include results from a simulation study in which we compare the mean squared errors (MSE) of the two alternatives. We conclude with a brief discussion in Section 5.

2. The finite population mixed model with auxiliary information and response error

We consider a finite population as a collection of N identifiable units, labeled $s = 1, \dots, N$, using a notation similar to that employed in Stanek, Singer & Lencina (2004). In particular, we assume that two fixed values, namely, the latent value y_s and the auxiliary variable x_s are associated with unit s . When unit s is selected, the auxiliary variable x_s is observed along with the response which is equal to the latent value, y_s , plus response error.

We formalize the process of simple random sampling by introducing a finite population model according to which any permutation of units in the population can be selected with equal probability $[(N!)^{-1}]$. We assign a new label, $i = 1, \dots, N$, to each unit according to its position in the permutation. We represent the latent value and the auxiliary variable associated to the unit in the i th position in the permutation respectively by Y_i and X_i . Note that Y_i and X_i are random variables because we do not know what unit will occupy the i th position in the permutation. On the other hand, y_s and x_s are fixed values, since they are associated to the unit labeled s . We may then represent a random permutation by a random matrix

$$(\mathbf{Y} \ \mathbf{X}) = \mathbf{U} (\mathbf{y} \ \mathbf{x}) \quad (2)$$

where $\mathbf{Y} = (Y_1, \dots, Y_i, \dots, Y_N)'$, $\mathbf{X} = (X_1, \dots, X_i, \dots, X_N)'$, $\mathbf{y} = (y_1, \dots, y_s, \dots, y_N)'$, $\mathbf{x} = (x_1, \dots, x_s, \dots, x_N)'$, $\mathbf{U} = (\mathbf{U}_1, \dots, \mathbf{U}_i, \dots, \mathbf{U}_N)'$ with $\mathbf{U}_i = (U_{i1}, \dots, U_{is}, \dots, U_{iN})'$ and U_{is} is an indicator random variable that takes on a value of one if unit s is selected in position i in the permutation, and zero otherwise. For example, let $N = 3$ then

$$\begin{pmatrix} Y_1 & X_1 \\ Y_2 & X_2 \\ Y_3 & X_3 \end{pmatrix} = \begin{pmatrix} U_{11} & U_{12} & U_{13} \\ U_{21} & U_{22} & U_{23} \\ U_{31} & U_{32} & U_{33} \end{pmatrix} \begin{pmatrix} y_1 & x_1 \\ y_2 & x_2 \\ y_3 & x_3 \end{pmatrix}. \text{ Suppose a realization}$$

$\mathbf{u} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ of $\mathbf{U}$ then $\begin{pmatrix} Y_1 & X_1 \\ Y_2 & X_2 \\ Y_3 & X_3 \end{pmatrix}_{\mathbf{U}=\mathbf{u}} = \begin{pmatrix} y_2 & x_2 \\ y_1 & x_1 \\ y_3 & x_3 \end{pmatrix}$. Letting the subscript P index expectation and variance with respect to permutation of units, we have $\mathbb{E}_P(\mathbf{U}) = N^{-1}\mathbf{J}_N$ and $\mathbb{V}_P[\text{vec}(\mathbf{U})] = (N-1)^{-1} \times (\mathbf{P}_N \otimes \mathbf{P}_N)$, for any $a \in \mathbb{Z}^+$, $\mathbf{J}_a = \mathbf{1}_a \mathbf{1}_a'$, $\mathbf{1}_a$ is an $a \times 1$ column vector with all elements equal to 1, and for any $b \neq 0$, $\mathbf{P}_{a,b} = \mathbf{I}_a - b^{-1} \mathbf{J}_a$, $\mathbf{P}_a = \mathbf{P}_{a,a}$, $\mathbf{I}_a$ is an $a \times a$ identity matrix, and vec and $\otimes$ denote the vec operator and the Kronecker product,

respectively, [see Harville (1997), for example]. For instance, let $a = 3$ and $b = 4$, so

$$J_3 = \mathbf{1}_3 \mathbf{1}_3^t = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} \text{ and } P_{3,4} = \begin{pmatrix} 1-1/4 & -1/4 & -1/4 \\ -1/4 & 1-1/4 & -1/4 \\ -1/4 & -1/4 & 1-1/4 \end{pmatrix}. \text{ Then it follows that}$$

$$\mathbb{E}_P \{ \text{vec} [(\mathbf{Y} \quad \mathbf{X})] \} = (\mathbf{I}_2 \otimes \mathbf{1}_N) \boldsymbol{\mu} \quad (3)$$

and

$$\mathbb{V}_P \{ \text{vec} [(\mathbf{Y} \quad \mathbf{X})] \} = \Sigma \otimes P_N,$$

where $\boldsymbol{\mu} = (\mu_y, \mu_x)'$, $\mu_y = N^{-1} \sum_{s=1}^N y_s$, $\mu_x = N^{-1} \sum_{s=1}^N x_s$ and

$$\Sigma = \begin{pmatrix} \sigma_y^2 & \sigma_{yx} \\ \sigma_{yx} & \sigma_x^2 \end{pmatrix}, \quad (4)$$

with $\sigma_y^2 = (N-1)^{-1} \sum_{s=1}^N (y_s - \mu_y)^2$, $\sigma_x^2 = (N-1)^{-1} \sum_{s=1}^N (x_s - \mu_x)^2$ and

$$\sigma_{yx} = (N-1)^{-1} \sum_{s=1}^N (y_s - \mu_y)(x_s - \mu_x).$$

We include response error associated to the unit s (not to position i) by adding the random vector $\mathbf{E} = (E_1, \dots, E_s, \dots, E_N)'$, where $\mathbb{E}_s(E_s) = 0$, $\mathbb{V}_R(E_s) = \sigma_s^2$ and $\mathbb{C}_{\mathbb{D}\mathbb{V}_R}(E_s, E'_s) = 0$ for $s \neq s'$, to the vector $\mathbf{y} = (y_1, \dots, y_s, \dots, y_N)'$ and considering the model

$$(\mathbf{Y}^* \quad \mathbf{X}) = \mathbf{U} (\mathbf{y} \quad \mathbf{x}) + \mathbf{U} (\mathbf{E} \quad \mathbf{0}_{N \times 1}), \quad (5)$$

where $\mathbf{Y}^* = (Y_1^*, \dots, Y_i^*, \dots, Y_N^*)'$. The subscript R indexes expectation and variance with respect to the response error distribution. Defining $\delta_{sy} = y_s - \mu_y$, so that $y_s = \mu_y + \delta_{sy}$, for position i in a permutation, we have

$$Y_i^* = \sum_{s=1}^N U_{is} (y_s + E_s) = \mu_y + d_{iy} + \sum_{s=1}^N U_{is} E_s, \quad (6)$$

where $d_{iy} = \sum_{s=1}^N U_{is} \delta_{sy}$ is a random unit effect, i.e., d_{iy} is the effect of the unit selected in i position with respect to μ_y . As in Li, Stanek III, and Singer (2012), we concatenate $\mathbf{Y}^*$ and $\mathbf{X}$ through the vec operator, and rewrite model (5) as

$$\mathbf{Z}^* = \text{vec} [(\mathbf{Y}^* \quad \mathbf{X})] = (\mathbf{I}_2 \otimes \mathbf{1}_N) \boldsymbol{\mu} + \mathbf{d} + \mathbf{E}^* \quad (7)$$

where $\mathbf{d} = (\mathbf{d}'_y, \mathbf{d}'_x)'$, $\mathbf{d}_y = (d_{1y}, \dots, d_{iy}, \dots, d_{Ny})'$, $\mathbf{d}_x = (d_{1x}, \dots, d_{ix}, \dots, d_{Nx})'$,
 $d_{ix} = \sum_{s=1}^N U_{is} \delta_{sx}$, $\delta_{sx} = x_s - \mu_x$, $\mathbf{E}^* = \begin{pmatrix} \mathbf{E}_y^* & \mathbf{0}'_{1 \times N} \end{pmatrix}'$, and $\mathbf{E}_y^* = \mathbf{Y}^* - \mathbf{U} \mathbf{y} = \mathbf{U} \mathbf{E}$. In
 Appendix A we show that

$$\mathbb{E}_{PR}(\mathbf{Z}^*) = (\mathbf{I}_2 \otimes \mathbf{1}_N) \boldsymbol{\mu}$$

and

$$\mathbb{V}_{PR}(\mathbf{Z}^*) = \sum \otimes \mathbf{P}_N + \begin{pmatrix} 1 & 0 \end{pmatrix}' \begin{pmatrix} 1 & 0 \end{pmatrix} \otimes \bar{\sigma}^2 \mathbf{I}_N$$

with $\bar{\sigma}^2 = N^{-1} \sum_{s=1}^N \sigma_s^2$ and $\mathbb{V}_R(E_s) = \sigma_s^2$. $\bar{\sigma}^2$ is called by Biemer (2009) as “simple response variance”.²

3. Target random variables and optimal estimators

In general, our interest is to estimate (predict) linear combinations of the form

$$T = \sum_{i=1}^N c_i Y_i, \quad (8)$$

where the c_i are known constants attached to positions in the permutation³. For instance, to specify the population total, we use $c_i = 1$ for all i and to specify the population mean, we employ $c_i = N^{-1}$ for all i .

To estimate T in (8) under the finite population mixed model (7) based on a simple random sample obtained without replacement, we consider linear estimators obtained via the following steps.

- i. We assume that the value of x_s is known for each of the N units⁴, and pre-multiply (7) by

$$\mathbf{R} = \begin{pmatrix} \mathbf{I}_N & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_N \end{pmatrix}$$

² $\bar{\sigma}^2 = N^{-1} \sum_{s=1}^N \sigma_s^2$ can be written as $\bar{\sigma}^2 = N^{-1} \sum_{i=1}^N \sigma_i^2$ where $\sigma_i^2 = \sum_{s=1}^N U_{is} \sigma_s^2$.

³ Depending on the choice of c_i , the quantity T may represent a parameter or a random effect. Note that T can be written as $T = \mathbf{c}'(\mu_y \mathbf{1}_N + \mathbf{d}_y)$ where $\mathbf{c} = (c_1, \dots, c_i, \dots, c_N)'$, and μ_y and $\mathbf{d}_y$ are defined in (3) and (7), respectively, so the target will be a parameter if $\mathbf{c}' \mathbf{d}_y = 0$, which occurs only if $\mathbf{c}$ is proportional to $\mathbf{1}_N$. For other values of $\mathbf{c}$, the target will be a random variable. We will use the term estimator in both cases.

⁴ For example, in the 1997 DANE monthly sample of retail trade in Colombia, the number of employees (auxiliary variable) was known for the 1705 businesses that constituted the target population. The sample contained 621 businesses. For details, see DANE (2009).

to eliminate μ_x . For the resulting model,

$$RZ^* = \begin{pmatrix} \mathbf{1}_N & \mathbf{0} \\ \mathbf{0} & \mathbf{0}_{N \times 1} \end{pmatrix} \mu + \mathbf{d} + \mathbf{E}^*,$$

where μ is defined in (3), and $\mathbf{d}$ and $\mathbf{E}^*$ are defined in (7). Since $RZ^* = \mu_y [(1 \ 0)' \otimes \mathbf{1}_N] + \mathbf{d} + \mathbf{E}^*$, it follows that $\mathbb{E}_{PR}(RZ^*) = \mu_y [(1 \ 0)' \otimes \mathbf{1}_N]$ and $\mathbb{V}_{PR}(RZ^*) = \mathbb{V}_{PR}(Z^*)$.

- ii. We partition $(Y^* \ X)$ into the sample (which may be taken as the first n rows, without loss of generality) and the remainder (the last $N - n$ rows). This step explicitly formalizes the process of simple random sampling. Letting

$$Z_I^* = [Y_1^*, \dots, Y_n^*, (X_1 - \mu_x), \dots, (X_n - \mu_x)]'$$

and

$$Z_{II}^* = [Y_{n+1}^*, \dots, Y_N^*, (X_{n+1} - \mu_x), \dots, (X_N - \mu_x)]',$$

it follows that the expected values and the variances of Z_I^* and Z_{II}^* are given by

$$\begin{aligned} \mathbb{E}_{PR}(Z_I^*) &= \mu_y [(1 \ 0)' \otimes \mathbf{1}_n], \\ \mathbb{E}_{PR}(Z_{II}^*) &= \mu_y [(1 \ 0)' \otimes \mathbf{1}_{N-n}], \\ \mathbb{V}_{PR}(Z_I^*) &= \mathbf{V}_I^* = \sum \otimes \mathbf{P}_{n,N} + (1 \ 0)' (1 \ 0) \otimes \sigma^2 \mathbf{I}_n, \\ \mathbb{V}_{PR}(Z_{II}^*) &= \mathbf{V}_{II}^* = \sum \otimes \mathbf{P}_{N-n,N} + (1 \ 0)' (1 \ 0) \otimes \sigma^2 \mathbf{I}_{N-n}, \\ \mathbb{Cov}_{PR}(Z_I^*, Z_{II}^*) &= \mathbf{V}_{I,II}^* = -N^{-1} \sum \otimes \mathbf{J}_{n \times (N-n)}, \end{aligned}$$

with $\mathbf{J}_{a \times b} = \mathbf{1}_a \mathbf{1}_b'$ for any $a \in \mathbb{Z}^+$ and $b \in \mathbb{Z}^+$.

We may write the target parameter as $T = \sum_{i=1}^n c_i Y_i + \sum_{i=n+1}^N c_i Y_i$ and since $\sum_{i=1}^n c_i Y_i$ is observed directly (when there is no response error) or indirectly (via $\sum_{i=1}^n c_i Y_i^*$ when there is response error), in order to estimate T , we have to estimate the second term, $\sum_{i=n+1}^N c_i Y_i$.

- iii. We minimize the expected mean squared error (MSE) of the linear estimator subject to an unbiased constraint (see Appendix B). This results in

$$\begin{aligned} \hat{T}^* &= N\bar{c}\hat{\alpha} + nN(\sigma_{yx}/\sigma_x^2)(\bar{X}_I - \mu_x)(\bar{c}_I - \bar{c})/(N-n) \\ &+ k^* \left(\sum_{i=1}^n c_i Y_i^* - n\bar{c}_I \bar{Y}_i^* \right) + (1-k^*) (\sigma_{yx}/\sigma_x^2) \left(\sum_{i=1}^n c_i X_i - n\bar{c}_I \mu_x \right) \\ &- n\bar{c}_I (1-k^*) (\sigma_{yx}/\sigma_x^2) (\bar{X}_I - \mu_x) \end{aligned} \quad (9)$$

with μ_x defined in (3), σ_x^2 , σ_y^2 and σ_{yx} defined in (4) and Y_i^* , in (6). We also define:

- $\bar{c} = N^{-1} \sum_{i=1}^N c_i$, $\bar{c}_I = n^{-1} \sum_{i=1}^n c_i$ with c_i defined in (8),
- $\hat{\alpha} = \bar{Y}_I^* + (\sigma_{yx}/\sigma_x^2)(\mu_x - \bar{X}_I)$,
- $k^* = \sigma_y^2(1 - \rho_{yx}^2)/[\bar{\sigma}^2 + \sigma_y^2(1 - \rho_{yx}^2)]$ with σ^2 defined in (7),
- $\rho_{yx} = \sigma_{yx}/(\sigma_y \sigma_x)$, $|\rho_{yx}| \leq 1$ denotes the correlation coefficient between y and x ,
- $\bar{Y}_I^* = n^{-1} \sum_{i=1}^n Y_i^*$ and $\bar{X}_I = n^{-1} \sum_{i=1}^n X_i$.

The first and the second terms in (9) relate to adjustments due to auxiliary information, while the remaining terms refer to adjustments due to response error (through k^*). Also, the estimator T^* can be written as the sum of an estimator of the portion of the target based on the sample, and an estimator of the portion of the target based on the remainder [see Appendix B, expression (16)].

The variance of the estimator T^* (see Appendix B) is

$$\mathbb{V}_{PR}(\hat{T}^* - T) = \sigma_y^2(1 - \rho_{yx}^2) \left[\frac{nD^2}{k^*} + (1 - k^*) \sum_{i=1}^n c_i^2 + (N - n) \bar{c}_{II}^2 \right] + \sigma_y^2 c_{II}^2, \quad (10)$$

where $D = (1/f - 1)c_{II} + (1 - k^*)c_I$ with $f = n/N$, the sampling fraction, and

$\bar{c}_{II} = (N - n)^{-1} \times \sum_{i=n+1}^N c_i$, and $c_{II}^2 = \sum_{i=n+1}^N (c_i - \bar{c}_{II})^2$. Note that to use (9) and (10), it is necessary to know σ_{yx} , σ_y and σ . These values are unknown in practice. We propose to use empirical estimates

$$\begin{aligned} \hat{\sigma}_{yx} &= (n-1)^{-1} \sum_{i=1}^n (\bar{Y}_i^* - \bar{Y}_I^*) (X_i - \bar{X}_I), \\ \hat{\sigma}_y^2 &= (n-1)^{-1} \sum_{i=1}^n (\bar{Y}_i^* - \bar{Y}_I^*)^2 \end{aligned}$$

and

$$\hat{\sigma}^2 = n^{-1} \sum_{i=1}^n \hat{\sigma}_i^2, \text{ with } \hat{\sigma}_i^2 = (m_i - 1)^{-1} \sum_{j=1}^{m_i} (Y_{ij}^* - \bar{Y}_i^*)^2,$$

where Y_{ij}^* , $j = 1, 2, \dots, m_i$, are a set of response measurements of the unit selected in i

position, $\bar{Y}_i^* = m_i^{-1} \sum_{j=1}^{m_i} Y_{ij}^*$ and $\bar{Y}_I^* = n^{-1} \sum_{i=1}^n \bar{Y}_i^*$.

To illustrate the use of T^* , consider the estimation of the population mean response, using (9) with $c_i = N^{-1}$ for all i . The resulting estimator is

$$\hat{T}^* = \bar{Y}_I^* + (\sigma_{yx}/\sigma_x^2) (\mu_x - \bar{X}_I)$$

$$\mathbb{V}_{PR}(\hat{T}^* - T) = n^{-1} [\sigma_y^2(1 - \rho_{yx}^2)(1 - f) + \bar{\sigma}^2] \quad (11)$$

Note that the shrinkage constant $k^* = \sigma_y^2(1 - \rho_{yx}^2) / [\bar{\sigma}^2 + \sigma_y^2(1 - \rho_{yx}^2)]$ does not appear in (11) because it is canceled out in the derivation. Simplifications of (11) may be obtained by considering the following assumptions for the variance and covariance terms

- a. when $\sigma_{yx} = 0$ and $\sigma^2 = 0$ (in other words, there are no auxiliary variables but there is response error),

$$\hat{T}^* = \bar{Y}_I^* \quad \text{and} \quad \mathbb{V}_{PR}(\hat{T}^* - T) = n^{-1} [\sigma_y^2(1 - f) + \bar{\sigma}^2];$$

- b. when $\sigma_{yx} \neq 0$ and $\sigma^2 = 0$ (in other words, there is an available auxiliary variable but there is no response error),

$$\hat{T} = \bar{Y}_I + (\sigma_{yx}/\sigma_x^2) (\mu_x - \bar{X}_I), \text{ where } \bar{Y}_I = n^{-1} \sum_{i=1}^n Y_i, \text{ and } \mathbb{V}_P(\hat{T} - T) = n^{-1} \sigma_y^2(1 - \rho_{yx}^2)(1 - f);$$

- c. when $\sigma_{yx} = 0$ and $\sigma^2 \neq 0$ (in other words, there are neither auxiliary variables nor response error),

$$\hat{T} = \bar{Y}_I \quad \text{and} \quad \mathbb{V}_P(\hat{T} - T) = n^{-1} \sigma_y^2(1 - f)$$

The estimator in a) is identical to the one considered in Bolfarine and Zacks (1992, p.151) under the location error-in-variables (superpopulation model) or to the one proposed by Särndal, Swensson, and Wretman (1992, p.605) under the simple measurement model, see also Biemer (2009)[p.283]. Case b) corresponds to the estimator given by Li, Stanek III, and Singer (2012), which also coincides with the estimator presented in Wu and Sitter (2001). Case c) corresponds to the π -estimator developed in Särndal, Swensson, and Wretman (1992, p.68).

Another example of the use of $\hat{T}^*$ is the estimation of the regression coefficient B , note that (1) may be written as (8) by taking $c_i = 1, i = 1, \dots, N$, and replacing $(y_s + E_s)$ by $[(y_s + E_s)w_s]$ where $w_s = (x_s - \mu_x) / [(N-1)\sigma_x^2], s = 1, \dots, N$. Then, the estimator is

$$\hat{B}^* = N\hat{T}^*, \quad (12)$$

where $\hat{T}^*$ is defined in (11). Then we may consider an empirical version of $\hat{B}^*$ given by

$$\hat{B}_e^* = N \left[\bar{Y}_I^* + (\hat{\sigma}_{yx}/\hat{\sigma}_x^2) (\mu_x - \bar{X}_I) \right] \quad (13)$$

where $\hat{\sigma}_x^2 = (n-1)^{-1} \sum_{i=1}^n (X_i - \bar{X}_I)^2$.

When $\sigma_{yx} = 0$, the estimator B^* reduces to

$$\bar{B}^* = N \bar{Y}_I^*. \quad (14)$$

4. Examples

For the example presented in the Introduction, the population consisted of the students admitted in 2004. Scores on different disciplines with a mathematical content were recorded for a simple random sample of size $n = 107$. The data are schematically displayed in Table 1. Mean corresponds to the average of the scores on the courses (Y_{i1}^* to Y_{i7}^*) actually attended by each student. $Y_i^* = Y_{i2}^*$ NA means that the student did not attend that course, so they are not used to evaluate Mean and $\hat{\sigma}_i^2$.

From the data, we obtain $\hat{\sigma}_x^2 = 25.10$ and $X_I = 20.93$. Although it is known that the FUVEST scores have a symmetric distribution, for confidential reasons we do not have access to the population values for the auxiliary variables⁵. We considered different values for μ_x and σ_x to compute the weights $W_i = (X_i - \mu_x) / [(N-1) \sigma_x^2]$, noting that using wrong values for such parameters may lead to biased estimates of T (9). We also replaced the observed response Y_i^* by the weighted response $Y_i^\bullet = Y_i^* W_i$ so that $\bar{Y}_I^*$ was replaced by $\bar{Y}_I^\bullet$ and used $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \hat{\sigma}_i^2$ as an estimate of σ^2 , the simple response variance, to compute $\hat{B}_e^*$ and $\hat{V}_{PR}(\hat{B}_e^* - B)$. The results are displayed in Table 2.

We observe that $\hat{V}_{PR}(\hat{B}_e^* - B)$ increases as $|\mu_x - X_I|$ increases, and σ_x^2 decrease. This suggests using an auxiliary variable with symmetric distribution, if it is available.

Alternatively, the corresponding ordinary least squares estimate is $B_{OLS} = 0.074$ with $\hat{V}(B_{OLS}) = 0.0004$ (the π -variance estimator)⁶. A difference between the estimated variances of the two estimators is that B_{OLS} takes no response error variance into account. Another difference is that the π -variance estimator is obtained via a Taylor linearization technique and our variance estimator is not. If we assume that the response error variance is null, it follows that $\hat{V}_P(\hat{B}_e^* - B) = 0.0005$.

⁵Our model requires the values for the auxiliary variable should be known for the entire population. However in practice it may be difficult to access them and the user needs to have approximations to the unknown μ_x and σ_x , usually obtained from some external source.

⁶The π -variance estimator of B_{OLS} under simple random sampling without replacement is given in Särndal, Swensson, and Wretman (1992, p. 196) by

$$\hat{V}(\hat{B}_{OLS}) = \frac{(1 - \frac{n}{N}) \frac{n}{n-1} \sum_{s \in \mathcal{S}} (x_s - \bar{x})^2 e_s^2}{\left[\sum_{s \in \mathcal{S}} (x_s - \bar{x})^2 \right]^2}$$

with $x = n^{-1} \sum_{s \in \mathcal{S}} x_s$, $e_s = y_s - \bar{y} - B_{OLS}(x_s - x)$ and $\bar{y} = n^{-1} \sum_{s \in \mathcal{S}} y_s$.

For illustrative purposes, we also consider a hypothetical population of size $N = 5$ and assume that the response error can take only two possible equally likely values given by plus or minus σ_s . We plan to select a simple random sample of size $n = 3$. In Table 3, we show the population values of the auxiliary variable (x_s), of the latent values (y_s), as well as the possible responses ($y_s \pm \sigma_s$) and the standard errors (σ_s) of E_s . The corresponding (population) scatterplot and the “least squares regression line” are presented in Figure 1.

Substituting y_s by $y_s^\bullet = y_s w_s$ (weighted latent value) where $w_s = (x_s - \mu_x) / \left[(N-1) \sigma_x^2 \right]$, $s = 1, \dots, N$, we obtain $\sigma^2 = 0.1305$, $\sigma_x^2 = 15.3$, $\sigma_{yx} / \sigma_x^2 = 0.0622$, $\mu_x = 5.4$, $\sigma_y^2 = 0.0708$ and $\nu_{yx} = \sigma_{yx} / (\sigma_y \sigma_x) = 0.9141$. We use the notation ν_{yx} instead of ρ_{yx} in order to distinguish the unweighted correlation coefficient (ρ_{yx}) from the weighted correlation coefficient (ν_{yx}). The population parameter of interest is $B = 0.4395$.

The results for the 960 ($= 5! \times 2^3$) possible permutations are schematically displayed in Table 4 along with the average of the estimates [$\hat{B}^*$ in (12), $\hat{B}_e^*$ in (13) and $\hat{B}^*$ in (14)], the relative contribution of the bias to the MSE [$= 100 \times (\text{estimate} - B)^2 / \text{MSE}$] and the MSE. We also include the ordinary least squares estimator, B_{OLS} . The results illustrate the optimal properties of B^* : it is unbiased and has a smaller MSE than the competing biased estimators $\hat{B}_e^*$ and B_{OLS} .

However, taking into account that to evaluate $\hat{B}^*$ it is necessary to know σ_{yx} , σ_y and σ_x , an interesting alternative is $\hat{B}_e^*$ in terms of MSE with respect to B_{OLS} .

As $\hat{B}_e^*$ and B_{OLS} show different results (when $\mu_x \neq \bar{X}_I$ and/or $\sigma_x^2 \neq \hat{\sigma}_x^2$) and as the proposed estimator B^* given in (12) depends on the covariance between the auxiliary variable (x_s) and the weighted latent values ($y_s^\bullet$), and the $y_s^\bullet$ depend on the latent values (y_s), we evaluate the effect of the coefficient of correlation between the auxiliary variable and the latent values, ρ_{yx} , on the MSE of B^* , of $\hat{B}_e^*$, and of B_{OLS} , through a small simulation study. We considered $n = (10, 30 \text{ or } 60)$ and $\sigma = (0, 0.25, 0.75, 1.5, 3, 5 \text{ or } 10)$. For each setting, we generate populations of size $N = 2000$ and different values of the correlation coefficient ($\rho_{yx} > 0$) between the latent values, y_s , and x_s , from

- i. a bivariate normal distribution, where $\mathbb{E}(Y) = \mathbb{E}(X) = 0$, $\mathbb{V}(Y) = \mathbb{V}(X) = 1$ and $\text{Cov}(X, Y) = \rho_{yx}$.
- ii. a bivariate finite distribution, asymmetrical for x , constructed by taking a bivariate normal distribution, with $\mathbb{E}(Y) = \mathbb{E}(X) = 0$, $\mathbb{V}(Y) = \mathbb{V}(X) = 1$ and $\text{Cov}(X, Y) = \rho_{yx}$, and replacing the generated values of x_s by $x_{s0} = (x_s / \sigma_x^2) e^{(x_s / \sigma_x^2)}$, limiting the range of ρ_{yx} to the interval 0 to 0.6.

The R code used in the computations is available from <http://www.docentes.unal.edu.co/lgonzalezg/docs/>

The results, displayed in Figures 2 and 3, show that

- i. for bivariate normal data [Figure 2], the MSE of $\hat{B}_{OLS}$ decreases as ρ_{yx} increases whereas, for bivariate data asymmetrical for x [Figure 3], the MSE of $\hat{B}_{OLS}$ increases as ρ_{yx} increases, and
- ii. for bivariate data asymmetrical for x [Figure 3], the MSE of $\hat{B}^*$ and $\hat{B}_e^*$ are smaller than the MSE of $\hat{B}_{OLS}$; however, for bivariate normal data [Figure 2], the MSE of $\hat{B}_{OLS}$ is smaller for large values of ρ_{yx} and for small values of σ .
- iii. If a user is using a response variable with high simple response variance or an auxiliary variable with asymmetric distribution, $\hat{B}_e^*$ is a good alternative with respect to $\hat{B}_{OLS}$, in particular to small samples.

Finally, we note that in general, along with the latent value y_s several auxiliary variables, $x_{1s}, \dots, x_{Qs}$, are associated to unit s . The extension of (9) to this setting is straightforward.

In this case, the estimator of T is given by

$$\hat{T}^* = N\bar{c}\hat{\alpha} + [nN(\bar{c}_I - \bar{c})/(N-n)]\bar{\mathbf{d}}_x' \boldsymbol{\gamma} + k^* \left(\sum_{i=1}^n c_i Y_i^* - n\bar{c}_I \bar{Y}_I^* \right) + (1-k^*) \left[\mathbf{c}_I' (\mathbf{d}_{1,I} \dots \mathbf{d}_{q,I} \dots \mathbf{d}_{Q,I}) - n\bar{c}_I \bar{\mathbf{d}}_x' \right] \boldsymbol{\gamma},$$

with Y_i^* defined in (6), c_i defined in (8), $\bar{c}_I$, $\bar{Y}_I^*$ and k^* (substituting ρ_{yx} by $\rho_{yx}^{(Q)}$) defined in (9).

- $\alpha = \bar{Y}_I^* - \bar{\mathbf{d}}_x' \boldsymbol{\gamma}$
- $\bar{\mathbf{d}}_x = (\bar{d}_{1x} \dots \bar{d}_{qx} \dots \bar{d}_{Qx})'$ with $\bar{d}_{qx} = X_{qI} - \mu_{xq}$, $\bar{X}_{qI} = n^{-1} \sum_{i=1}^n X_{qi}$ and $\mu_{xq} = N^{-1} \sum_{s=1}^N x_{qs}$,
- $\boldsymbol{\gamma} = (\gamma_1 \dots \gamma_q \dots \gamma_Q)' = \boldsymbol{\sigma}_{xx}^{-1} \boldsymbol{\sigma}_{yx}$, $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_y^2 & \boldsymbol{\sigma}_{yx}' \\ \boldsymbol{\sigma}_{yx} & \boldsymbol{\sigma}_{xx} \end{pmatrix}$, $\boldsymbol{\sigma}_{yx} = (\sigma_{yx_1} \dots \sigma_{yx_Q})'$, $\boldsymbol{\sigma}_{xx} = \begin{pmatrix} \sigma_{x_1}^2 & \dots & \sigma_{x_1 x_Q} \\ \vdots & \ddots & \vdots \\ \sigma_{x_Q x_1} & \dots & \sigma_{x_Q}^2 \end{pmatrix}$ and $\sigma_{x_q x_{q'}}'$ as defined in (4);
- $\mathbf{c}_I = (c_1 \dots c_i \dots c_n)'$,
- $\mathbf{d}_{q,I} = ([X_{q1} - \mu_{xq}] \dots [X_{qi} - \mu_{xq}] \dots [X_{qn} - \mu_{xq}])'$, $q = 1, \dots, Q$,
- $\rho_{yx}^{(Q)} = \sqrt{\boldsymbol{\sigma}_{yx}' \boldsymbol{\sigma}_{xx}^{-1} \boldsymbol{\sigma}_{yx} / \sigma_y^2}$ is the multiple correlation coefficient between the response y and the auxiliary variables $\mathbf{x}_1, \dots, \mathbf{x}_q, \dots, \mathbf{x}_Q$, $\mathbf{x}_q = (x_{q1} \dots x_{qN})'$, [see Morrison (1983), for example]. We assume that $|\rho_{yx}^{(Q)}| \neq 1$.

Finally, $\text{Var}_{SR}(\hat{T}^* - T)$ is given by (10) with the substitution of ρ_{yx} by $\rho_{yx}^{(Q)}$.

5. Discussion

The best linear unbiased estimator (9) is appropriate for a wide variety of problems. The required assumptions are minimal and used in many stochastic models. Target quantities may correspond to the mean, to regression coefficients, or, in general, to linear combinations of individual latent values. Here we only considered cases where the population means of the auxiliary variables are known, and that their values are observed without error. Cases where this does not hold need further investigation.

Of particular interest is the estimation of the regression coefficient (1) which corresponds to a weighted mean, with weights $w_s = (x_s - \mu_x) / [(N-1) \sigma_x^2]$ applied to the individual responses. The sample mean $[B^*$ in (14)] is an intuitive estimator of B and it corresponds to the best linear unbiased estimator when no additional auxiliary information is taken into account. However, by including the auxiliary variable in the model, we may obtain a more accurate estimator. The increase in accuracy is due to an adjustment (for example, post-stratification) to account for the difference between the sample mean of the auxiliary variable and its population mean. Since the auxiliary variable and the weighted response are correlated, using an adjustment proportional to the deviation between the sample mean response and the auxiliary variable population mean can lead to an improvement in the accuracy of the resulting estimator.

From (11) it follows that the variance of B^* decreases as $|v_{yx}|$ increases. Further improvement may be obtained with the inclusion of additional auxiliary variables that increase the multiple correlation with the response.

A more complicated picture emerges when comparing a standard estimator, B_{OLS} with B^* , or with the empirical estimator $\hat{B}_e^*$ (13). First, we note that for symmetric response distributions, when (σ_{yx}/σ_x^2) in (9) is known, B_{OLS} may outperform B^* if the multiple correlation coefficient is large. This result holds, even though B^* is the BLUP. A possible explanation relies on the differences between the model-based and the sampling/response error frameworks. In the sampling/response error framework, B_{OLS} is not a linear estimator, since values of the auxiliary variables for a sample are random (and depend on the realized units). When the distribution of the auxiliary variable is skewed, however, the performance of B^* or even of the empirical estimator, $\hat{B}_e^*$, appears to improve dramatically. This suggests that further investigation is needed to provide practical advice for the selection of appropriate estimators.

Next, we note that (9) can be used to estimate realized random effects and extend the results of Li, Stanek III, and Singer (2012) to settings where there is response error, or those of Stanek & Singer (2004) to settings that include adjustment for auxiliary variables. The most common situation occurs when weights are all taken equal to one, and c_i in (8) is one for i and zero for $i > n$. In such a context for a single auxiliary variable, the BLUP of the latent value for the j -th selected unit is given by

$$\hat{T}^* = \bar{Y}_I^* + k^*(Y_j^* - \bar{Y}_I^*) + (\sigma_{yx}/\sigma_x^2)(1 - k^*)(X_{1j} - \bar{X}_{1I}).$$

The shrinkage constant, $k^* = \sigma_y^2(1 - \rho_{yx}^2) / [\bar{\sigma}^2 + \sigma_y^2(1 - \rho_{yx}^2)]$, is attenuated by the correlation coefficient between the response and the auxiliary variable. In contrast, the shrinkage constant for the least squares solution, $1 - k^* = \bar{\sigma}^2 / [\bar{\sigma}^2 + \sigma_y^2(1 - \rho_{yx}^2)]$, increases with the correlation coefficient.

We finalize by highlighting the small value of the correlation coefficient between the FUVES Math score and the (latent) ability in Mathematics obtained in the example. This could be viewed as an indication that either the FUVES Math exam does not select students with ability in Mathematics or that the scores obtained in the different disciplines do not reflect this.

Extension to samples with stratified sampling using simple random sampling without replacement in all strata (STSI design) may be considered with an appropriate definition of the U matrix in (2), i.e., letting $U^* = \text{diag}\{U^{(1)}, \dots, U^{(H)}\}$ where a superscript (h) refers to stratum h , and $U^{(h)} = (U_{h1}, \dots, U_{hN_h})'$, with $U_{hi} = (U_{hi1}, \dots, U_{hiN_h})'$ and U_{his} is an indicator random variable that takes on a value of one if unit s in stratum h , $h = 1, \dots, H$, is selected in position i (in stratum h), and zero otherwise. Derivation of the corresponding estimator and of its variance involves rather tedious matrix algebra, but is feasible. Extension to simple random (without replacement) cluster sampling (SIC design) requires replacing y_s and x_s in (2) by the respective cluster totals, but we have not obtained the corresponding results. Extensions to unequal selection probabilities are more complicated and need further investigation and might require jackknife or bootstrap methods.

FUVES: Mathematics Score

References

- Baker, FB. Item Response Theory - Parameters Estimation Techniques. Marcel Dekker, Inc; New York: 1992.
- Biemer, P. Measurement errors in surveys. In: Pfeffermann, D.; Rao, CR., et al., editors. Handbook of statistics, Vol 29, Sample Surveys: Design, Methods and Applications. Vol. 12. Elsevier; Oxford: 2009. p. 281-315.
- Bolfarine, H.; Zacks, S. Prediction Theory for Finite Populations. Springer-Verlag; New York: 1992.
- Bolfarine H, Zacks S, Elian SN, Rodrigues J. Optimal prediction of the finite population regression coefficient. *Sankhyâ B*. 1994; 56:1-10.
- DANE. Metodología muestra mensual de comercio al por menor. 2009. Retrieved August 29, 2013, from: <http://www.dane.gov.co/files/investigaciones/fichas/MMCM.pdf>
- Fuller WA. Regression analysis for sample survey. *Sankhyâ C*. 1975; 37:117-132.
- Hambleton RK, Cook LL. Latent trait models and their use in the analysis of educational test data. *Journal of Educational Measurement*. 1997; 14:75-96.
- Harville, DA. Matrix Algebra from a Statistician's Perspective. Springer; New York: 1997.
- Holt D, Smith TMF, Winter PD. Regression analysis of data from complex surveys. *Journal of the Royal Statistical Society, Series A*. 1980; 143:474-487.
- Kish L, Frankel MR. Inference from complex samples (with discussion). *Journal of the Royal Statistical Society, Series B*. 1974; 36:1-37.

- Li W, Stanek EJ III, Singer JM. Design-based random permutation models with auxiliary information. *Statistics: A Journal of Theoretical and Applied Statistics*. 2012; 46:663–671.
- Little RJ. To model or not to model? competing modes of inference for finite population sampling. *Journal of the American Statistical Association*. 2004; 99:546–556.
- Morrison, DF. *Applied Linear Statistical Methods*. Prentice-Hall; New Jersey: 1983.
- Neter, J.; Kutner, MH.; Nachtsheim, CJ.; Wasserman, W. *Applied Linear Statistical Models*. WCB/McGraw-Hill; Boston: 1996.
- Pfeffermann D, Holmes DJ. Robustness considerations in the choice of a method of inference for regression analysis of survey data. *Journal of the Royal Statistical Society, Series A*. 1985; 148:268–278.
- Pfeffermann D, Smith TMF. Regression models for grouped populations in cross-section surveys. *International Statistical Review*. 1985; 53:37–59.
- Sen, A.; Srivastava, M. *Regression Analysis. Theory, Methods, and Applications*. Springer-Verlag; New York: 1990.
- Särndal, CE.; Swensson, B.; Wretman, J. *Model Assisted Survey Sampling*. Springer-Verlag; New York: 1992.
- Stanek EJ III, Singer JM. Predicting random effects from finite population clustered samples with response error. *Journal of the American Statistical Association*. 2004; 99:1119–1130.
- Stanek EJ III, Singer JM, Lencina VB. A unified approach to estimation and prediction under simple random sampling. *Journal of Statistical Planning and Inference*. 2004; 121:325–338.
- Wu C, Sitter RR. A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*. 2001; 96:185–193.

Appendix A: Derivation of (Z^*) and (Z^*) in (7)

$$\begin{aligned}\mathbb{E}_{PR}(Z^*) &= \mathbb{E}_P \left\{ \text{vec}[\mathbf{U}(\mathbf{y} \ \mathbf{x})] \right\} + \mathbb{E}_{PR} \left\{ \text{vec}[\mathbf{U}(\mathbf{E} \ \mathbf{0}_{N \times 1})] \right\} \\ &= \mathbb{E}_P \left\{ \text{vec}[\mathbf{U}(\mathbf{y} \ \mathbf{x})] \right\} = (\mathbf{I}_2 \otimes \mathbf{1}_N) \boldsymbol{\mu}\end{aligned}$$

because $\mathbb{E}_{PR}(U_{is}E_s) = \mathbb{E}_P(U_{is}) \mathbb{E}_R(E_s) = 0$, and

$$\begin{aligned}\mathbb{V}_{PR}(Z^*) &= \mathbb{V}_{PR} \left\{ \text{vec}[\mathbf{U}(\mathbf{y} \ \mathbf{x})] + \text{vec}[\mathbf{U}(\mathbf{E} \ \mathbf{0})] \right\} \\ &= \mathbb{E}_P \left(\mathbb{V}_{R|S} \left\{ \text{vec}[\mathbf{U}(\mathbf{y} \ \mathbf{x})] + \text{vec}[\mathbf{U}(\mathbf{E} \ \mathbf{0})] \right\} \right) + \mathbb{V}_P \left(\mathbb{E}_{R|S} \left\{ \text{vec}[\mathbf{U}(\mathbf{y} \ \mathbf{x})] + \text{vec}[\mathbf{U}(\mathbf{E} \ \mathbf{0})] \right\} \right),\end{aligned}$$

where

$$\mathbb{V}_{R|S} \left\{ \text{vec}[\mathbf{U}(\mathbf{y} \ \mathbf{x})] + \text{vec}[\mathbf{U}(\mathbf{E} \ \mathbf{0})] \right\} = (\mathbf{1} \ 0)' (\mathbf{1} \ 0) \otimes \left[\mathbf{U} \left(\bigoplus_{s=1}^N \sigma_s^2 \right) \mathbf{U}' \right]$$

and

$$\mathbb{E}_{R|S} \left\{ \text{vec}[\mathbf{U}(\mathbf{y} \ \mathbf{x})] + \text{vec}[\mathbf{U}(\mathbf{E} \ \mathbf{0})] \right\} = \text{vec}[\mathbf{U}(\mathbf{y} \ \mathbf{x})].$$

Then

$$\mathbb{V}_{PR}(\mathbf{Z}^*) = \begin{pmatrix} 1 & 0 \end{pmatrix}' \begin{pmatrix} 1 & 0 \end{pmatrix} \otimes \bar{\sigma}^2 \mathbf{I}_N + \sum \otimes \mathbf{P}_N$$

$$\text{with } \bar{\sigma}^2 = \mathbb{E}_P \left(\sum_{s=1}^N \sigma_s^2 U_{is} \right) = N^{-1} \sum_{s=1}^N \sigma_s^2.$$

Appendix B: Development of the linear estimator of T (8) and its variance under model (7)

In order to obtain the best linear unbiased estimator of $T = \sum_{i=1}^N c_i Y_i$ under model (7), we consider the following three conditions:

- i. it must be a linear function of the sample random variables, in other words,

$$\hat{T}^* = (\mathbf{g}'_I + \mathbf{a}') \mathbf{Z}_I^* \text{ where } \mathbf{g}_I = \begin{pmatrix} \mathbf{c}'_I & \mathbf{0}'_{1 \times n} \end{pmatrix}' \text{ with}$$

$$\mathbf{c}_I = (c_1 \dots c_i \dots c_n)', \quad (15)$$

c_i defined in (8), $\mathbf{Z}_I^* = \begin{pmatrix} [\mathbf{Y}_I^*]' & [X_1 - \mu_x] & \dots & [X_i - \mu_x] & \dots & [X_n - \mu_x] \end{pmatrix}'$ with $\mathbf{Y}_I^* = (Y_1^* \dots Y_i^* \dots Y_n^*)'$, and X_i, μ_x and Y_i^* defined in (2), (3) and (6), respectively,

- ii. it must be unbiased, in other words, $\mathbb{E}_{PR}[(\mathbf{g}'_I + \mathbf{a}') \mathbf{Z}_I^*] = \mathbb{E}_P(\mathbf{g}'_I \mathbf{Z}_I + \mathbf{g}'_{II} \mathbf{Z}_{II})$ where

$$\begin{aligned} \mathbf{g}_{II} &= \begin{pmatrix} \mathbf{c}'_{II} & \mathbf{0}'_{1 \times (N-n)} \end{pmatrix}', \mathbf{c}_{II} = (c_{n+1} \dots c_i \dots c_N)', \\ \mathbf{Z}_I &= \begin{pmatrix} Y_1 \dots Y_n & [X_1 - \mu_x] & \dots & [X_i - \mu_x] & \dots & [X_n - \mu_x] \end{pmatrix}', \\ \mathbf{Z}_{II} &= \begin{pmatrix} Y_{n+1} \dots Y_N & [X_{n+1} - \mu_x] & \dots & [X_i - \mu_x] & \dots & [X_N - \mu_x] \end{pmatrix}', \end{aligned}$$

and Y_i defined in (2), and

- iii. it must minimize $\mathbb{V}_{PR}(\hat{T}^* - T)$.

The estimator resulting in

$$\hat{T}^* = \mathbf{g}'_I [\mathbf{X}_I^o \hat{\alpha} + \mathbf{V}_I [\mathbf{V}_I^*]^{-1} (\mathbf{Z}_I^* - \mathbf{X}_I^o \hat{\alpha})] + \mathbf{g}'_{II} [\mathbf{X}_{II}^o \hat{\alpha} + \mathbf{V}'_{II} [\mathbf{V}_I^*]^{-1} (\mathbf{Z}_I^* - \mathbf{X}_I^o \hat{\alpha})] \quad (16)$$

where $\hat{\alpha} = (\mathbf{X}_I^o' [\mathbf{V}_I^*]^{-1} \mathbf{X}_I^o)^{-1} \mathbf{X}_I^o' [\mathbf{V}_I^*]^{-1} \mathbf{Z}_I^*$, as in Stanek & Singer (2004), where

$$\begin{aligned}
\mathbf{X}_I^o &= \begin{pmatrix} 1 & 0 \end{pmatrix}' \otimes \mathbf{1}_n, \\
\mathbf{X}_{II}^o &= \begin{pmatrix} 1 & 0 \end{pmatrix}' \otimes \mathbf{1}_{N-n}, \\
\mathbf{V}_1 &= \mathbb{V}_P(\mathbf{Z}_I) = \sum \otimes \mathbf{P}_{n,N}, \\
\mathbf{V}_I^* &= \sum \otimes \mathbf{P}_{n,N} + \begin{pmatrix} 1 & 0 \end{pmatrix}' \begin{pmatrix} 1 & 0 \end{pmatrix} \otimes \bar{\sigma} \mathbf{I}_n, \\
\mathbf{V}_{I,II} &= \mathbb{Cov}_S(\mathbf{Z}_I, \mathbf{Z}_{II}) = \mathbb{Cov}_{PR}(\mathbf{Z}_I^*, \mathbf{Z}_{II}^*) = \mathbb{Cov}_{PR}(\mathbf{Z}_I^*, \mathbf{Z}_{II}^*) = \mathbf{V}_{I,II}^* \\
&= N^{-1} \sum \otimes \mathbf{J}_{n \times (N-n)}.
\end{aligned}$$

The first term in the expression (16) is a estimator of the portion of the target corresponding to the sample. The second term is the portion of the target corresponding to the remainder.

In (16) for $\mathbf{g}_I' [\mathbf{X}_I^o \hat{\alpha} + \mathbf{V}_I [\mathbf{V}_I^*]^{-1} (\mathbf{Z}_I^* - \mathbf{X}_I^o \hat{\alpha})]$, the second term, in other words,

$\mathbf{V}_I [\mathbf{V}_I^*]^{-1} (\mathbf{Z}_I^* - \mathbf{X}_I^o \hat{\alpha})$, is an adjustment due to response error.

We can group terms in the estimator as

$$\hat{T}^* = (\mathbf{g}_I' \mathbf{X}_I^o + \mathbf{g}_{II}' \mathbf{X}_{II}^o) \hat{\alpha} + (\mathbf{g}_I' \mathbf{V}_I + \mathbf{g}_{II}' \mathbf{V}_{I,II}') [\mathbf{V}_I^*]^{-1} (\mathbf{Z}_I^* - \mathbf{X}_I^o \hat{\alpha})$$

so that the contribution of the sample and remaining random variables to the estimator are considered in each term. We interpret the second term as the adjustment due to response error.

Simplifying, it follows that $\mathbf{g}_I' \mathbf{X}_I^o + \mathbf{g}_{II}' \mathbf{X}_{II}^o = N\bar{c}$ with $\bar{c}$ defined in (9),

$\mathbf{g}_I' \mathbf{V}_I + \mathbf{g}_{II}' \mathbf{V}_{I,II}' = \begin{pmatrix} \sigma_y^2 & \sigma_{yx} \end{pmatrix} \otimes (\bar{c}_I' - \bar{c} \mathbf{1}_n')$ with σ_y^2 and σ_{yx} defined in (4), and $\bar{c}_I$ defined in (15),

$$[\mathbf{V}_I^*]^{-1} = \sum^{-1} \otimes \mathbf{P}_{n,n-N} - \sum^{-1} \begin{pmatrix} 1 & 0 \end{pmatrix}' \begin{pmatrix} 1 & 0 \end{pmatrix} \sum^{-1} \otimes \mathbf{F},$$

where Σ is defined in (4), $\sum^{-1} = \begin{pmatrix} \sigma_{11}^o & \sigma_{12}^o \\ \sigma_{12}^o & \sigma_{22}^o \end{pmatrix}$ and

$\mathbf{F} = (\sigma_{11}^o)^{-1} (1 - k^*) \times \mathbf{P}_{n,n-N} \mathbf{P}_{n,h} \mathbf{P}_{n,n-N}$ with $h^{-1} = (1 - k^*) / (N - nk^*)$, and

$$\begin{aligned}
\hat{\alpha} &= [\sigma_{11}^o n N k^* / (N - nk^*)]^{-1} [n N k^* / (N - nk^*)] \left[\sigma_{11}^o \bar{Y}_I^* + \sigma_{12}^o (\bar{X}_I - \mu_x) \right] \\
&= \bar{Y}_I^* + (\sigma_{12}^o / \sigma_{11}^o) (\bar{X}_I - \mu_x),
\end{aligned}$$

where μ_x is defined in (3), and k^* , $\bar{Y}_I^*$ and $\bar{X}_I$ in (9). Using this results,

$$\begin{aligned}\hat{T}^* &= N\bar{c}\hat{\alpha} + \left[\begin{pmatrix} \sigma_y^2 & \sigma_{yx} \\ \sigma_{yx} & \sigma_x^2 \end{pmatrix} \otimes (\mathbf{c}'_I - \bar{c}\mathbf{1}'_n) \right] [\mathbf{V}_I^*]^{-1} (\mathbf{Z}_I^* - \mathbf{X}_I^o \hat{\alpha}) \\ &= N\bar{c}\hat{\alpha} + k^* \mathbf{c}'_I \mathbf{P}_n \mathbf{Y}_I^* + (1-k^*) (\sigma_{yx}/\sigma_x^2) \mathbf{c}'_I (\mathbf{X}_I - \mu_x \mathbf{1}_n) + [\sigma_{yx}/\sigma_x^2] [\bar{\mathbf{X}}_I - \mu_x] [nN(\bar{c}_I - \bar{c})/(N-n) - n\bar{c}_I(1-k^*)],\end{aligned}$$

where σ_x^2 is defined in (4) and $\mathbf{X}_I = (X_1 \dots X_i \dots X_n)'$. Simplifying, the resulting expression is given in (9). This expression for the estimator is valid for any $\mathbf{c} = (c_1 \dots c_i \dots c_N)'$.

The variance of the estimator given in (9) is

$$\mathbb{V}_{PR}(\hat{T}^* - T) = \hat{\mathbf{a}}' \mathbf{V}_I^* \hat{\mathbf{a}} + \mathbf{g}'_I (\mathbf{V}_I^* - \mathbf{V}_I) \mathbf{g}_I + 2\hat{\mathbf{a}}' (\mathbf{V}_I^* - \mathbf{V}_I) \mathbf{g}_I - 2\hat{\mathbf{a}}' \mathbf{V}_{I,II} \mathbf{g}_{II} + \mathbf{g}'_{II} \mathbf{V}_{II} \mathbf{g}_{II} \quad (17)$$

where

$$\begin{aligned}\hat{\mathbf{a}} &= \begin{pmatrix} 1 \\ -\sigma_{yx}/\sigma_x^2 \end{pmatrix} \otimes [D\mathbf{1}_n - (1-k^*) \mathbf{c}_I] + \bar{c}_{II} \begin{pmatrix} 0 \\ -\sigma_{yx}/\sigma_x^2 \end{pmatrix} \otimes \mathbf{1}_n, \\ \mathbf{V}_{II} &= \mathbb{V}_P(\mathbf{Z}_{II}) = \sum \otimes \mathbf{P}_{N-n,N}, \\ \hat{\mathbf{a}}' \mathbf{V}_I^* \hat{\mathbf{a}} &= \sigma_y^2 (1-\rho_{yx}^2) \sum_{i=1}^n \frac{1}{k^*} [D - (1-k^*)c_i]^2 + \frac{(N-n)\sigma_y^2 \bar{c}_{II}^2}{N} [n - N(1-\rho_{yx}^2)], \\ \mathbf{g}'_I (\mathbf{V}_I^* - \mathbf{V}_I) \mathbf{g}_I &= \sigma_y^2 (1-\rho_{yx}^2) \left(\frac{1-k^*}{k^*} \right) \sum_{i=1}^n c_i^2, \\ \hat{\mathbf{a}}' (\mathbf{V}_I^* - \mathbf{V}_I) \mathbf{g}_I &= \sigma_y^2 (1-\rho_{yx}^2) \left(\frac{1-k^*}{k^*} \right) \sum_{i=1}^n c_i [D - (1-k^*)c_i], \\ \hat{\mathbf{a}}' \mathbf{V}_{I,II} \mathbf{g}_{II} &= -\frac{(N-n)\sigma_y^2 \bar{c}_{II}^2}{N} [N(1-\rho_{yx}^2) - n], \\ \mathbf{g}'_{II} \mathbf{V}_{II} \mathbf{g}_{II} &= (N-n-1)\sigma_y^2 \sigma_{c_{II}}^2 + \frac{n(N-n)}{N} \bar{c}_{II}^2 \sigma_y^2,\end{aligned}$$

with ρ_{yx} and k^* defined in (9), D , c_{II} and $\sigma_{c_{II}}^2$ defined in (10), and $\mathbf{c}_I$ in (15). The resulting expression is given in (10).

Appendix C: Interpretation of the measure of association between the auxiliary variable and the latent values

To interpret B , we consider the (nonstochastic) linear model

$$y_s = A + Bx_s + e_s, \quad s=1, \dots, N, \quad (18)$$

where $A = \mu_y - B\mu_x$, B , μ_y and μ_x defined in (1), and $e_s = y_s - \mu_y - B(x_s - \mu_x)$. Under the descriptive (frequentist, randomization or design-based) approach discussed in Kish and Frankel (1974), A and B may be respectively interpreted as the values of the vector $(A \ B)'$ that minimize $\sum_{s=1}^N (y_s - a - bx_s)^2$, and, Särndal, Swensson, and Wretman (1992, p.191) mention that “By way of interpretation, $(A \ B)'$ is the coefficient vector of the best fitting plane of the type $y = A + Bx$ ”. They also state that “ B measures the number of units of y per unit of x in the finite population” and comment that “An analyst interested in a regression

slope, B , implicitly considers that a linear regression model is consistent with the finite population scatter”.

On the other hand, under the so-called superpopulation model (model-based) approach, it is assumed that the finite population is a random sample from an infinite ‘superpopulation’ so that, the observations on the dependent and independent variables follow the classical linear model [see Pfeiffermann and Holmes (1985) and Little (2004), for example]. Under this approach the simple linear regression model assumes that y_s^* is random and may be expressed as a sum of a deterministic element $\alpha + \beta x_s$ and a random element ε_s , in other words,

$$y_s^* = \alpha + \beta x_s + \varepsilon_s, \quad s=1, \dots, N, \quad (19)$$

where α and β are the unknown regression coefficients. We use the superscript * in order to distinguish the random nature of y_s^* (induced by ε_s) from that of y_s in (18) which is fixed. For the random vector $\boldsymbol{\varepsilon} = (\varepsilon_1 \dots \varepsilon_s \dots \varepsilon_N)'$, we assume $\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0}_{N \times 1}$ and a positive-definite covariance matrix $\mathbf{V}$; therefore $\mathbb{E}_{\Psi}(y_s^*/x_s) = \alpha + \beta x_s$, $s = 1, \dots, N$, where the expectation, $\mathbb{E}_{\Psi}(\cdot)$, is taken with respect to the superpopulation linear regression model parameterized by $\Psi = (\alpha, \beta, \mathbf{V})$. For uncorrelated and homoscedastic errors, $\mathbf{V} = \sigma^2 \mathbf{I}_N$, $\sigma^2 > 0$, where $\mathbf{I}_N$ denotes the identity matrix of dimension N . Consequently, under model-based approach, Neter, Kutner, Nachtsheim, and Wasserman (1996, p.12) interpret the parameter β as “the change in the mean of the probability distribution of y^* per unit increase in x ”, and, Sen and Srivastava (1990, p.6) state that “for every unit increase in x , y^* increases by β ”.

For finite populations, the above interpretation may not be appropriate because the mean of y is a constant which does not depend on x . Otherwise, we may write

$$B = \sum_{\substack{s=1 \\ x_s \neq \mu_x}}^N \frac{(x_s - \mu_x)^2}{\sum_{s=1}^N (x_s - \mu_x)^2} \times \frac{(y_s - \mu_y)}{(x_s - \mu_x)} = \sum_{\substack{s=1 \\ \tilde{w}_s \neq 0}}^N \tilde{w}_s b_s,$$

where $\tilde{w}_s = (x_s - \mu_x)^2 / \sum_{s=1}^N (x_s - \mu_x)^2$ and $b_s = (y_s - \mu_y) / (x_s - \mu_x)$ when $x_s \neq \mu_x$, and x_s, y_s, μ_x and μ_y are defined in (1). Then B may be viewed as a weighted average of the N slopes, b_s , of the lines passing through each (x_s, y_s) and (μ_x, μ_y) . These slopes might be interpreted as measures of variation of y per unit variation of x , so that B may be interpreted as a weighted average measure of association, with the weights, $\tilde{w}_s$, depending on the distance between x_s and μ_x .

Figure 1. Scatterplot and least squares regression line ($y = 1.4265 + 0.4395x$) for the hypothetical example.

Figure 2.

MSE ($\times 10^{-2}$) of $\hat{B}^*$, $\hat{B}_e^*$ and $\hat{B}_{OLS}$ for different values of ρ_{yx} , n (the first number within parentheses) and σ (the second number within parentheses). Bivariate normal data.

Figure 3.

MSE ($\times 10^{-2}$) of $\hat{B}^*$, $\hat{B}_e^*$ and $\hat{B}_{OLS}$ for different values of ρ_{yx} , n (the first number within parentheses) and σ (the second number within parentheses). Asymmetrical data for x .

Table 1

Schematic representation of data for the FUVEST example

Student	FUVEST	Y_{i1}^{**}	Y_{i2}^{**}	Y_{i3}^{**}	Y_{i4}^{**}	Y_{i5}^{**}	Y_{i6}^{**}	Y_{i7}^{**}	Mean	σ_i^2
1	35	8.4	6.9	NA	8.0	NA	5.0	NA	7.075	2.316
2	34	6.8	9.8	NA	NA	NA	NA	2.3	6.300	14.250
3	24	6.7	5.4	NA	NA	6.8	2.7	7.0	5.720	3.247
4	21	7.2	5.7	5.2	NA	NA	5.6	NA	5.925	0.769
5	30	9.7	5.1	NA	NA	8.6	5.3	NA	7.175	5.409
6	24	6.2	5.5	NA	NA	5.2	6.4	NA	5.825	0.322
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
102	15	7.5	6.7	NA	7.6	NA	6.9	NA	7.175	0.196
103	21	6.6	5.6	NA	NA	NA	6.2	NA	6.133	0.253
104	15	6.4	7.7	NA	8.9	NA	NA	8.0	7.750	1.070
105	22	6.3	5.1	NA	6.7	NA	5.5	NA	5.900	0.533
106	20	5.9	5.1	7.6	3.0	NA	4.7	5.0	5.217	2.278
107	14	7.8	3.6	NA	7.5	NA	5.5	NA	6.100	3.820

Y_{i1}^{**} =Introduction to probability I

Y_{i3}^{**} =Linear Algebra I

Y_{i5}^{**} =Linear Algebra for Computer Science

Y_{i7}^{**} =Introduction to Probability II (2005)

Y_{i2}^{**} = Calculus I

Y_{i4}^{**} =Linear Algebra

Y_{i6}^{**} =Introduction to Probability II (2004)

Table 2

$\hat{B}_e^* \left(\hat{\Psi}_{PR} \left[\hat{B}_e^* - B \right] \times 10^{-3} \right)$ considering different values of μ_x and σ_x^2 in the FUVES example

μ_x	σ_x^2					
	15	20	25.10	30	35	40
15	-0.205 (9.8)	-0.154 (5.5)	-0.122 (3.5)	-0.102 (2.5)	-0.088 (1.8)	-0.077 (1.4)
17	-0.055 (6.8)	-0.041 (3.8)	-0.033 (2.4)	-0.028 (1.7)	-0.024 (1.2)	-0.021 (1.0)
19	0.055 (4.9)	0.041 (2.8)	0.033 (1.8)	0.027 (1.2)	0.023 (0.9)	0.021 (0.7)
20.93	0.123 (4.3)	0.093 (2.4)	0.074 (1.5)	0.062 (1.1)	0.053 (0.8)	0.046 (0.6)
23	0.156 (4.9)	0.117 (2.7)	0.093 (1.7)	0.078 (1.2)	0.067 (0.9)	0.058 (0.7)
25	0.147 (6.7)	0.110 (3.8)	0.088 (2.4)	0.073 (1.7)	0.063 (1.2)	0.055 (0.9)

Table 3

The hypothetical finite population

Unit labels (s)	x_s	y_s	$y_s \pm \sigma_s$	σ_s
1	1	1	-2 or 4	3
2	3	2.5	1.5 or 3.5	1
3	4	5	-2 or 12	7
4	10	6.5	-3.5 or 16.5	10
5	9	4	2 or 6	2

$$\rho_{yx} = 0.80$$

Table 4

Finite population mixed model estimates of B with known or unknown σ_{yx} and σ_x^2 , ordinary least square estimate, relative contribution of the bias to the MSE and average MSE with $N = 5$ and $n = 3$.

Sample	X_1	X_2	X_3	Y_1^*	Y_2^*	Y_3^*	B^*	$\hat{B}_e^*$	B^{*-}	$\hat{B}_{OLS}$
1	1	3	4	-2	1.5	-2	1.068	-0.363	0.218	0.250
2	1	3	4	-2	1.5	12	0.534	-2.147	-0.316	4.250
...	...	...	...	...	...	...	...	...	...	...
959	4	10	3	12	16.5	1.5	1.429	1.233	1.511	1.657
960	4	10	3	12	16.5	3.5	1.298	1.092	1.381	1.471
Average							0.4395	0.2789	0.4395	0.4722
100× bias ² /MSE							0.0%	2.2%	0.0%	0.1%
MSE							1.1264	1.1712	1.3236	1.6619