BEST LINEAR UNBIASED LATENT VALUES PREDICTORS FOR FINITE POPULATION LINEAR MODELS WITH DIFFERENT ERROR SOURCES

Germán MORENO[1]
Julio da Motta SINGER[2]
Edward J. STANEK III[3]

▪ ABSTRACT: We develop best linear unbiased predictors (BLUP) of the latent values of labeled sample units selected from a finite population when there are two distinct sources of measurement error: endogenous, exogenous or both. Usual target parameters are the population mean, the latent values associated to a labeled unit or the latent value of the unit that will appear in a given position in the sample. We show how both types of measurement errors affect the within unit covariance matrices and indicate how the finite population BLUP may be obtained via standard software packages employed to fit mixed models in situations with either heteroskedastic or homoskedastic exogenous and endogenous measurement errors.

▪ KEYWORDS: BLUP; covariance matrix; measurement error.

1 Introduction

Predicting the latent value (expected value) of a variable for a sample unit on which some measurements are made is a common problem in Applied Statistics. Sometimes, the response variable is subject to different sources of variability associated to measurement errors as indicated in Cochran (1977) or observation errors as termed by Sukhatme *et al.* (1984). Two sources of measurement errors can

[1] Universidad Industrial de Santander, Escuela de Matemáticas, Bucaramanga, Colombia. *In memorian*

[2] Universidade de São Paulo - USP, Departamento de Estatística, CEP:05508090, São Paulo, SP, Brasil. E-mail: *jmsinger@ime.usp.br*

[3] University of Massachusetts, Department of Public Health, Amherst, Massachusetts, USA. E-mail: *stanek@schoolph.umass.edu*

be identified. The first is related to the natural variability of the unit responses and is referred to as inherent variability in the terminology introduced by Buonaccorsi (2006) or response error by Särndal, Swensson and Wretman (1992). The second is associated with the measuring conditions and it corresponds to the variability of the measures around a fixed value (the latent value), produced by the measurement instruments or interviewers, for example. To clearly differentiate between the two types of measurement errors, we refer to the first as endogenous measurement errors and to the second, as exogenous measurement errors.

Endogenous measurement errors may occur even if the measuring is made with absolute precision (*i.e.*, with no exogenous measurement error). The monthly expenditure with food for a given family is an example; the expenditure may vary from month to month around a latent value, but can be measured without error. Measurement of an adult's height by different observers may serve as an example of a situation where there is only exogenous errors. The results of the daily measurement of a patient's cholesterol level may serve as an example of a situation where both endogenous and exogenous measurement errors are present.

As an example, we consider data for a subset of 13 participants in the project *Seasonal Variability of Blood Lipids, NHLBI, number R01-HL52745* (MERRIAM *et al.*, 1999). Data in this study were collected with the goal of identifying and quantifying factors that relate to seasonal changes in cholesterol. For each participant, triplicate measures of cholesterol were in collected in four quarters. In each quarter, the data were collected not necessarily by the same examiner. We reproduce part of the data in Table 1, that for illustrative effect, will represent our target population.

Table 1 - Example of data of the *Seasonal Variability of Blood Lipids*

Name	Data	Examiner	Cholesterol	Quarter
1	13/05/1996	CS	208,6	1
1	15/05/1996	CS	206,9	1
1	17/05/1996	CS	208,4	1
1	12/08/1996	SU	171,3	2
1	16/08/1996	SU	174,3	2
1	20/08/1996	SU	185,7	2
⋮	⋮	⋮	⋮	⋮
13	26/08/1996	KL	182,0	3
13	28/08/1996	KL	194,5	3
13	29/08/1996	KL	198,5	3
13	22/11/1996	SU	107,1	4
13	24/11/1996	SU	102,9	4
13	27/11/1996	SU	109,1	4

We let y_s denote the latent cholesterol level for the unit labeled s, $s = 1, \ldots, N$, *i.e.*, the expected value of the cholesterol level over 4 quarters and represent the corresponding endogenous measurement error variance by σ_s^2. The population mean cholesterol level is $\mu = N^{-1}\sum_{s=1}^{N} y_s$ and the population variance is $\gamma^2 = (N-1)^{-1}\sum_{s=1}^{N}(y_s-\mu)^2$.

We assume that the variability in the response introduced by the examiner is the exogenous measurement error. For unit labeled s, measured in quarter q by the j-th examiner, we represent the observed response by

$$Y_{sqj} = y_{sq} + \widetilde{W}_j, \tag{1}$$

where y_{sq} represents the latent level of cholesterol for unit s in quarter q and $\widetilde{W}_j$ represents exogenous measurement error, assumed to have mean zero and variance $\widetilde{\sigma}_j^2$. The question is how can we estimate the latent value y_s of unit s in the population.

Linear mixed models have been extensively used for such purposes in an infinite population setup as indicated in Goldberger (1962), Verbeke and Molenberghs (2000), McCulloch and Searle (2001), Diggle, Heagerty, Liang and Zeger (2002), Demidenko (2013), Fitzmaurice, Davidian, Verbeke and Molenberghs (2008), among others. The standard linear mixed model for the response from the i-th unit selected from a population can be represented as

$$Y_i = \mu + B_i + E_i, \quad i = 1, \ldots, n, \tag{2}$$

where μ is the population mean response, B_i is a random effect corresponding to the i-th selected unit, assumed to have mean zero and variance γ^2 and E_i is a measurement error, assumed to have mean zero and variance σ^2 (or σ_i^2, for heteroskedastic models).

What does E_i represent? The answer depends on the manner with which response error is associated with the realized units. If we assume that there is no exogenous errors, then E_i represents the inherent variability of the i-th selected unit response. Now, if assume that there is no variability in the selected unit's response and that all variability is due to the effect of measuring, then we can say that E_i is associated to the exogenous variability.

What happens with E_i when you have the two types of variability simultaneously? If W_i represents the endogenous measurement error and $\widetilde{W}_i$ the exogenous measurement error, then $E_i = f(W_i, \widetilde{W}_i)$. The standard linear mixed model (2) does not consider the distinction between the two sources of measurement errors. It also does not retain identifiability of the units in the population. Our objective is to clarify such issues in a finite population setup.

In Section 2, we describe the finite population mixed model with endogenous/exogenous measurement errors and derive optimal estimators or predictors using the expanded variable approach considered in Singer *et al.* (2012) along with the methodology employed in standard linear mixed models and we discuss the relationship between the predictors obtained under both approaches for different covariance structures. In Section 3, we compare latent value predictors obtained via finite population and standard linear mixed models. In Section 4, we analyse the cholesterol data described in the Introduction and indicate how the function `lme` in the statistical software package R may be employed to fit finite population linear mixed models in situations with either heteroskedastic or homoskedastic exogenous and endogenous errors. We conclude with a brief discussion in Section 5.

2 The finite population mixed model

We define a finite population as a collection of N identifiable units labeled $s = 1, \dots, N$. Let $\boldsymbol{y} = (y_1, \dots, y_N)^\top$ denote a vector for which the s-th element is the response latent value y_s associated with unit s. The population mean response is $\mu = N^{-1}\sum_{s=1}^{N} y_s$, and the population response variance is $\gamma^2 = N^{-1}\sum_{s=1}^{N}(y_s - \mu)^2 = N^{-1}\sum_{s=1}^{N} b_s^2$ where $b_s = y_s - \mu$. Note that in this setup, b_s is a constant and not a random effect.

Following Singer *et al.* (2012), we define the random permutation model as an ordered list of N random variables, where units are independently permuted. For each permutation, we assign a new label, $i = 1, \dots, N$ to the units according to their position in the permuted list, letting $\boldsymbol{Y} = (Y_1, \dots, Y_N)^\top$ denote the random vector of latent permuted values. Simple random sampling without replacement is introduced via a set of correlated indicator random variables, U_{is}, that take on a value of one with probability $1/N$ if unit s is selected in position i in the sample and zero otherwise. Then $\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{y}$, where

$$\boldsymbol{U} = \begin{pmatrix} U_{11} & U_{12} & \dots & U_{1N} \\ U_{21} & U_{22} & \dots & U_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ U_{N1} & U_{N2} & \dots & U_{NN} \end{pmatrix}.$$

Letting the subscript ξ_1 denote expectation with respect to the permutation distribution, it follows that

$$\left.\begin{array}{l}
\bullet \quad \mathbb{E}_{\xi_1}(U_{is}) = \dfrac{1}{N}, \; i = 1, \dots, N, s = 1, \dots, N, \\
\bullet \quad \mathbb{V}_{\xi_1}(U_{is}) = \dfrac{1}{N}\left(1 - \dfrac{1}{N}\right), \; i = 1, \dots, N, s = 1, \dots, N, \\
\bullet \quad \mathbb{COV}_{\xi_1}(U_{is}, U_{i^*s^*}) = -\dfrac{1}{N^2}, \; i = i^* \text{ and } s \neq s^*, \\
\bullet \quad \mathbb{CV}_{\xi_1}(U_{is}, U_{i^*s^*}) = -\dfrac{1}{N^2}, \; i \neq i^* \text{ and } s = s^*, \\
\bullet \quad \mathbb{COV}_{\xi_1}(U_{is}, U_{i^*s^*}) = \dfrac{1}{N^2(N-1)}, \; i \neq i^* \text{ and } s \neq s^*, \\
\bullet \quad \mathbb{E}_{\xi_1}\left[vec(\boldsymbol{U}^\top)\right] = \dfrac{1}{N}\left(\mathbf{1}_N \otimes \mathbf{1}_N\right), \\
\bullet \quad \mathbb{V}_{\xi_1}\left[vec(\boldsymbol{U}^\top)\right] = \dfrac{1}{N-1}\boldsymbol{P}_N \otimes \boldsymbol{P}_N,
\end{array}\right\} \tag{3}$$

where, $\boldsymbol{P}_N = \boldsymbol{I}_N - \frac{1}{N}\boldsymbol{J}_N$ with $\boldsymbol{J}_N = \boldsymbol{I}_N\boldsymbol{I}_N^\top$, $\boldsymbol{I}_N$ denotes an identity matrix of dimension N and $\mathbf{1}_N$ denotes an $N \times 1$ vector with all elements equal to one. Then, it follows that

$$\mathbb{E}_{\xi_1}(\boldsymbol{Y}) = \mathbf{1}_N\mu \;\text{ and }\; \mathbb{V}_{\xi_1}(\boldsymbol{Y}) = \gamma^2[\boldsymbol{I}_N - \frac{1}{N}\boldsymbol{J}_N]. \tag{4}$$

Suppose that a simple random sample without replacement is to be selected from the population. Without loss of generality, we let the sample (indexed by $i = 1, \ldots, n$) consist of the elements occupying the first $n \leq N$ positions in a permutation . If we assume that only one observation is made on the i-th selected unit and no measurement errors are considered, the model for the observable response Y_i in i-th position is

$$Y_i = \sum_{s=1}^{N} U_{is}y_s = \mu + B_i. \tag{5}$$

When endogenous measurement error W_s associated to unit s is present, the model for the observable response $\widetilde{Y}_i$ may be specified as

$$\widetilde{Y}_i = \sum_{s=1}^{N} U_{is}(y_s + W_s). \tag{6}$$

If, in addition, an exogenous measurement error is considered for the j-th measurent condition, the model for the observable response $\widetilde{Y}_i$ is

$$\widetilde{Y}_i = \sum_{s=1}^{N} U_{is}(y_s + W_s) + \widetilde{\widetilde{W}}_i. \tag{7}$$

Expression (7) may be written as

$$\begin{aligned} \widetilde{Y}_i &= \sum_{s=1}^{N} U_{is}(\mu + b_s + W_s) + \widetilde{\widetilde{W}}_i = \mu + \sum_{s=1}^{N} U_{is}b_s + \sum_{s=1}^{N} U_{is}W_s + \widetilde{\widetilde{W}}_i \qquad (8) \\ &= \mu + B_i + W_i^* + \widetilde{\widetilde{W}}_i, \end{aligned}$$

where $W_i^* = \sum_{s=1}^{N} U_{is}W_s$ denotes the endogenous measurement error associated to the i-th selected unit and $B_i = \sum_{s=1}^{N} U_{is}b_s$ denotes a random effect. Note that (8) has a similar expression as the standard linear mixed model (2), with the exception that the two sources of measurement error terms (endogenous and exogenous) are explicit in the former. The standard linear mixed model cannot distinguish these two sources of measurement error since the subscript i indexes simultaneously the position and the selected unit in the sample.

Since $\sum_{s=1}^{N} U_{is} = 1$ for all $i = 1, \ldots, N$, and in each row of $\boldsymbol{U}$, there exists a single value equal to 1, all the other being zero, it follows that $\widetilde{\widetilde{W}}_i = \sum_{s=1}^{N} U_{is}\widetilde{\widetilde{W}}_i$.

Then, when both endogenous and exogenous measurement errors are present, a model for response on the N positions in the permuted population is

$$\widetilde{\boldsymbol{Y}} = \left[\sum_{s=1}^{N} U_{1s}(y_s + W_s + \widetilde{W}_1), \ldots, \sum_{s=1}^{N} U_{Ns}(y_s + W_s + \widetilde{W}_N)\right]^{\top}.$$

Letting the subscript ξ_2 represent expectation with respect to the endogenous measurement error and subscript ξ_3 represent expectation with respect to the exogenous measurement error, we consider the following assumptions

$$\left.\begin{array}{l}
\bullet\ \ \mathbb{E}_{\xi_2}(W_s) = 0,\ s = 1, \ldots, N, \\
\bullet\ \ \mathbb{V}_{\xi_2}(W_s) = \sigma_s^2,\ s = 1, \ldots, N, \\
\bullet\ \ \mathbb{COV}_{\xi_2}(W_s, W_{s^*}) = 0,\ s, s^* = 1, \ldots, N, s \neq s^*, \\
\bullet\ \ \mathbb{E}_{\xi_3}(\widetilde{W}_i) = 0,\ i = 1, \ldots, N, \\
\bullet\ \ \mathbb{V}_{\xi_3}(\widetilde{W}_i) = \sigma_i^2,\ i = 1, \ldots, N, \\
\bullet\ \ \mathbb{COV}_{\xi_3}(\widetilde{W}_i, \widetilde{W}_{i^*}) = 0,\ i, i^* = 1, \ldots, N, i \neq i^*, \\
\bullet\ \ \mathbb{E}_{\xi_2\xi_3|\xi_1}(W_s + \widetilde{W}_i) = 0,\ s = 1, \ldots, N,\ i = 1, \ldots, N, \\
\bullet\ \ \mathbb{V}_{\xi_2\xi_3|\xi_1}(W_s + \widetilde{W}_i) = \sigma_s^2 + \sigma_i^2,\ s = 1, \ldots, N,\ i = 1, \ldots, N.
\end{array}\right\} \tag{9}$$

it follows that the expected value and variance of the random variable $\widetilde{\boldsymbol{Y}}$ are, respectively

$$\mathbb{E}_{\xi_1\xi_2\xi_3}(\widetilde{\boldsymbol{Y}}) = \mathbf{1}_N \mu \tag{10}$$

and

$$\mathbb{V}_{\xi_1\xi_2\xi_3}(\widetilde{\boldsymbol{Y}}) = \gamma^2 \boldsymbol{P}_N + \overline{\sigma}^2 \boldsymbol{I}_N + \bigoplus_{i=1}^{N} \sigma_i^2 \tag{11}$$

where $\overline{\sigma}^2 = N^{-1} \sum_{s=1}^{N} \sigma_s^2$ and $\bigoplus$ denotes the direct sum operator.

We are interested in developing an optimal linear unbiased predictor (or estimate) of target quantities of the form $P = \boldsymbol{g}^\top \boldsymbol{Y}$ where $\boldsymbol{g}$ is a vector of constants. For example,

i) if $\boldsymbol{g} = \mathbf{1}_N$, then $\boldsymbol{g}^\top \boldsymbol{Y} = \sum_{i=1}^{N} Y_i = \tau$, is the population total;

ii) if $\boldsymbol{g} = N^{-1}\mathbf{1}_N$, then $\boldsymbol{g}^\top \boldsymbol{Y} = N^{-1} \sum_{i=1}^{N} Y_i = \mu$, is the population mean;

iii) if $\boldsymbol{g} = \boldsymbol{e}_i$, with $\boldsymbol{e}_i$ denoting a vector with null elements except for the i-th which is equal to 1, then $\boldsymbol{g}^\top \boldsymbol{Y} = \mu + B_i$, the latent value of the unit in the i-th position of the random permutation.

Note that i) and ii) represent fixed values but iii) refers to a random variable. We are interested in predicting the random variable in iii). For such purpose, we follow the ideas of Singer *et al.* (2012) and consider a setup to develop the BLUP of the target quantity.

First, we represent a simple random sample without replacement by the first $n \leq N$ random variables in $\widetilde{\boldsymbol{Y}}$ and let the remaining $(N-n)$ random variables denote the responses of the non-sampled elements. Explicitly, we let $\widetilde{\boldsymbol{Y}} = [\widetilde{\boldsymbol{Y}}_S^\top, \widetilde{\boldsymbol{Y}}_R^\top]^\top$ and will express the predictor as a linear combination of the sample random variables, $\widetilde{\boldsymbol{Y}}_S$. To determine the coefficients of these random variables that lead to the optimal predictor, we specify an unbiasedness constraint and then minimize the expected mean squared error, subject to this constraint. This leads to the BLUP of the target.

Taking (3) and (9) into account, we have

$$\mathbb{E}_{\xi_1\xi_2\xi_3}\left[\begin{array}{c}\widetilde{\boldsymbol{Y}}_S\\ \widetilde{\boldsymbol{Y}}_R\end{array}\right] = \left[\begin{array}{c}\mathbf{1}_n\\ \mathbf{1}_{N-n}\end{array}\right]\mu \tag{12}$$

and

$$\mathbb{V}_{\xi_1\xi_2\xi_3}\left[\begin{array}{c}\widetilde{\boldsymbol{Y}}_S\\ \widetilde{\boldsymbol{Y}}_R\end{array}\right] = \left[\begin{array}{cc}\widetilde{\mathbb{V}}_S & \widetilde{\mathbb{V}}_{SR}\\ \widetilde{\mathbb{V}}_{SR}^\top & \widetilde{\mathbb{V}}_R\end{array}\right] = \gamma^2\left[\begin{array}{cc}\boldsymbol{I}_n - \frac{1}{N}\boldsymbol{J}_n & -\frac{1}{N}\boldsymbol{J}_{n\times(N-n)}\\ & \\ -\frac{1}{N}\boldsymbol{J}_{(N-n)\times n} & \boldsymbol{I}_{N-n} - \frac{1}{N}\boldsymbol{J}_{N-n}\end{array}\right] \tag{13}$$

$$+\,\overline{\sigma}^2\left[\begin{array}{cc}\boldsymbol{I}_n & \mathbf{0}_{n\times(N-n)}\\ & \\ \mathbf{0}_{(N-n)\times n} & \boldsymbol{I}_{N-n}\end{array}\right] + \left[\begin{array}{cc}\bigoplus_{i=1}^{n}\sigma_i^2 & \mathbf{0}_{n\times(N-n)}\\ & \\ \mathbf{0}_{(N-n)\times n} & \bigoplus_{i=n+1}^{N}\sigma_i^2\end{array}\right].$$

Now, we let $\boldsymbol{g}^\top = (\boldsymbol{g}_S^\top, \boldsymbol{g}_R^\top)$ so that the quantity to predict is $P = \boldsymbol{g}_S^\top\boldsymbol{Y}_S + \boldsymbol{g}_R^\top\boldsymbol{Y}_R$. The BLUP of P must satisfy the following criteria considered in Royall (1976), *i.e.*, it must:

- be a linear combination of the sample data: $\widehat{P} = \boldsymbol{c}^\top\widetilde{\boldsymbol{Y}}_S$
- be unbiased: $\mathbb{E}_{\xi_1\xi_2\xi_3}\left[\boldsymbol{c}^\top\widetilde{\boldsymbol{Y}}_S - (\boldsymbol{g}_S^\top\boldsymbol{Y}_S + \boldsymbol{g}_R^\top\boldsymbol{Y}_R)\right] = 0$
- have minimum MSE, $\mathbb{V}_{\xi_1\xi_2\xi_3}\left[\boldsymbol{c}^\top\widetilde{\boldsymbol{Y}}_S - (\boldsymbol{g}_S^\top\boldsymbol{Y}_S + \boldsymbol{g}_R^\top\boldsymbol{Y}_R)\right]$.

The unbiasedness constraint implies that $\boldsymbol{c}^\top\mathbb{E}(\widetilde{\boldsymbol{Y}}_S) = \boldsymbol{g}_S^\top\mathbb{E}(\boldsymbol{Y}_S) + \boldsymbol{g}_R^\top\mathbb{E}(\boldsymbol{Y}_R)$ which reduces to

$$\boldsymbol{c}^\top\mathbf{1}_n = \boldsymbol{g}^\top\mathbf{1}_N \tag{14}$$

given that $\boldsymbol{g}_S^\top\mathbb{E}(\boldsymbol{Y}_S) + \boldsymbol{g}_R^\top\mathbb{E}(\boldsymbol{Y}_R) = \boldsymbol{g}^\top\mathbf{1}_N\mu$ and from (12), $\mathbb{E}(\widetilde{\boldsymbol{Y}}_S) = \mathbf{1}_n\mu$.

Observing that

$$\mathbb{V}_{\xi_1\xi_2\xi_3}\left[\begin{array}{c}\widetilde{\boldsymbol{Y}}_S\\ \boldsymbol{Y}_S\\ \boldsymbol{Y}_R\end{array}\right] = \left[\begin{array}{ccc}\widetilde{\mathbb{V}}_S & \mathbb{V}_S & \mathbb{V}_{SR}\\ \mathbb{V}_S & \mathbb{V}_S & \mathbb{V}_{SR}\\ \mathbb{V}_{SR}^\top & \mathbb{V}_{SR}^\top & \mathbb{V}_R\end{array}\right]$$

and recalling (4) and (13), we obtain

$$\mathbb{V}_{\xi_1\xi_2\xi_2}(\widehat{P}-P) = \boldsymbol{c}^\top\widetilde{\mathbb{V}}_S\boldsymbol{c} + \boldsymbol{g}_S^\top\mathbb{V}_S\boldsymbol{g}_S + \boldsymbol{g}_R^\top\mathbb{V}_R\boldsymbol{g}_R - 2\boldsymbol{c}^\top\mathbb{V}_S\boldsymbol{g}_S - 2\boldsymbol{c}^\top\mathbb{V}_{SR}\boldsymbol{g}_R - 2\boldsymbol{g}_R^\top\mathbb{V}_{SR}^\top\boldsymbol{g}_S. \tag{15}$$

Therefore, using Lagrangian multipliers, we seek the value of $\boldsymbol{c}$ that will minimize

$$f(\boldsymbol{c}, \lambda) = \mathbb{V}_{\xi_1\xi_2\xi_2}(\widehat{P} - P) + \lambda(\boldsymbol{c}^\top \mathbf{1}_n - \boldsymbol{g}_R^\top \mathbf{1}_{N-n}).$$

Differentiating with respect to $\boldsymbol{c}$ and λ, setting these derivatives to zero and solving for $\boldsymbol{c}$ we obtain the BLUP of P as

$$\widehat{P} = \boldsymbol{g}_S^\top[\mathbf{1}_n\widehat{\mu} + \mathbb{V}_S\widetilde{\mathbb{V}}_S^{-1}(\widetilde{\boldsymbol{Y}}_S - \mathbf{1}_n\widehat{\mu})] + \boldsymbol{g}_R^\top[\mathbf{1}_{N-n}\widehat{\mu} + \mathbb{V}_{SR}^\top\widetilde{\mathbb{V}}_S^{-1}(\widetilde{\boldsymbol{Y}}_S - \mathbf{1}_n\widehat{\mu})] \tag{16}$$

with

$$\widehat{\mu} = (\mathbf{1}_n^\top\widetilde{\mathbb{V}}_S^{-1}\mathbf{1}_n)^{-1}\mathbf{1}_n^\top\widetilde{\mathbb{V}}_S^{-1}\widetilde{\boldsymbol{Y}}_S. \tag{17}$$

For details, see Singer *et al.* (2012).

In particular, to obtain the BLUP of the latent value $P_i = \mu + B_i$ associated to the i-th selected unit in the sample, first observe that

$$\begin{aligned}
\widetilde{V}_S &= \gamma^2\left[\boldsymbol{Y}_n - \frac{1}{N}\boldsymbol{J}_n\right] + \overline{\sigma}^2\boldsymbol{I}_n + \bigoplus_{i=1}^{n}\sigma_i^2 \\
V_S &= \gamma^2\left[\boldsymbol{Y}_n - \frac{1}{N}\boldsymbol{J}_n\right] \\
V_{SR} &= -\gamma^2\frac{1}{N}[\mathbf{1}_n\mathbf{1}_{N-n}^\top].
\end{aligned}$$

Given that

$$\widetilde{V}_S^{-1} = \bigoplus_{i=1}^{n}\left(\gamma^2 + \overline{\sigma}^2 + \sigma_i^2\right)^{-1} + \frac{\gamma^2}{N - \gamma^2 L}\boldsymbol{m}\boldsymbol{m}^\top,$$

where $L = \sum_{i=1}^{n}\left(\gamma^2 + \overline{\sigma}^2 + \sigma_i^2\right)^{-1}$ and $\boldsymbol{m}$ is an $n\times 1$ vector with the i-th component equal to $\left(\gamma^2 + \overline{\sigma}^2 + \sigma_i^2\right)^{-1}$, the remaining ones equal to zero, from (17), it follows that

$$\widehat{\mu} = \sum_{i=1}^{n} k_i\widetilde{Y}_i / \sum_{i=1}^{n} k_i,$$

where $k_i = \left(\gamma^2 + \overline{\sigma}^2 + \sigma_i^2\right)^{-1}$.

Then, (16) simplifies to

$$\widehat{P}_i = \widehat{\mu} + \frac{\gamma^2}{\gamma^2 + \overline{\sigma}^2 + \sigma_i^2}(\widetilde{Y}_i - \widehat{\mu}), \quad i \leq n \tag{18}$$

where $\gamma^2/(\gamma^2 + \overline{\sigma}^2 + \sigma_i^2)$ is a shrinkage constant.

When there are only endogenous measurement errors, the shrinkage constant is $\gamma^2/(\gamma^2 + \overline{\sigma}^2)$ and the BLUP is

$$\widehat{P}_i = \overline{\widetilde{Y}} + \frac{\gamma^2}{\gamma^2 + \overline{\sigma}^2}(\widetilde{Y}_i - \overline{\widetilde{Y}}), \quad i \leq n \tag{19}$$

with $\overline{\widetilde{Y}} = n^{-1}\sum_{i=1}^{n}\widetilde{Y}_i$.

When there are only exogenous measurement errors, the shrinkage constant is $\gamma^2/(\gamma^2+\sigma_i^2)$ and the BLUP is

$$\widehat{P}_i = \widehat{\mu} + \frac{\gamma^2}{\gamma^2+\sigma_i^2}(\widetilde{Y}_i - \widehat{\mu}), \quad i \leq n. \tag{20}$$

When neither measurement errors are present, the BLUP reduces to $\overline{Y}_i = n^{-1}\sum_{i=1}^{n} Y_i$. In practice, the variance components must be estimated, leading to the so called empirical BLUP.

3 A comparison of finite population and standard mixed models predictors

To clarify the effect of different sources of measurement errors in the prediction of latent effects under mixed models, we reproduce a simple example from Singer *et al.* (2012). For such purpose, we compare predictors of latent values of sampled units in the presence of endogenous heteroskedastic measurement errors. We consider a population of size $N = 3$ from which a sample of size $n = 2$ is selected. A single measurement of a response variable with two possible values (equal to the latent value ± the endogenous standard error) is obtained on each sampled unit. The population parameters are presented in Table 2. The idea is to compare the performance of the usual heteroskedastic linear mixed model BLUP, namely,

Table 2 - Parameters of a hypothetical finite population

	Latent	Variance	Weight	Shrinkage constant
Label	value	σ_i^2	k_i	$[\gamma^2/(\gamma^2+\sigma_i^2)]$
Alba	$y_1 = 10$	$\sigma_1^2 = 1$	$k_1 = 0.491$	$w_1 = 0.950$
Juliana	$y_1 = 3$	$\sigma_2^2 = 100$	$k_2 = 0.082$	$w_s = 0.160$
Laura	$y_1 = 2$	$\sigma_3^2 = 4$	$k_3 = 0.427$	$w_s = 0.826$
	$\mu = 5$			
	$\gamma^2 = 19$	$\overline{\sigma}^2 = 35$		$w = 0.352$

$$\widehat{Q}_i = \widehat{\mu} + \frac{\gamma^2}{\gamma^2+\sigma_i^2}(Y_i - \widehat{\mu}) \tag{21}$$

with that of the corresponding heteroskedastic finite population mixed model BLUP,

$$\widehat{P}_i = \overline{Y} + \frac{\gamma^2}{\gamma^2+\overline{\sigma}^2}(Y_i - \overline{Y}).$$

In Table 3 we present all the possible results for samples of size $n = 2$ along with the corresponding BLUP $\widehat{Q}_i$ and $\widehat{P}_i$ as well as their squared errors.

Table 3 - Possible results obtained with a sample of size $n = 2$ from the population described in Table 2 along with the corresponding BLUP ($\widehat{Q}_i$ and $\widehat{P}_i$) along with their respective squared errors, $(\widehat{Q}_i - y_s)^2$ or $(\widehat{P}_i - y_s)^2$

	Latent values		Observed values ($Y_i = y_{s_i} \pm s_i$)		BLUP				Squared error	
Sample	y_{s_1}	y_{s_2}	Y_1	Y_2	$\widehat{Q}_1$	$\widehat{Q}_2$	$\widehat{P}_1$	$\widehat{P}_2$	$\widehat{Q}$	$\widehat{P}$
Alba/Juliana	10	3	11	13	11.0	11.6	11.6	12.4	37.2	45.1
Alba/Juliana	10	3	11	-7	10.9	5.9	5.2	-1.2	4.7	20.4
Alba/Juliana	10	3	9	13	9.0	10.1	10.3	11.7	25.8	37.9
Alba/Juliana	10	3	9	-7	8.9	4.5	3.8	-1.8	1.7	30.7
Alba/Laura	10	2	11	4	10.8	4.7	8.7	6.3	3.9	9.9
Alba/Laura	10	2	11	0	10.7	1.0	7.4	3.6	0.8	4.5
Alba/Laura	10	2	9	4	8.9	4.5	7.4	5.6	3.7	10.0
Alba/Laura	10	2	9	0	8.8	0.8	6.1	2.9	1.4	8.1
Juliana/Alba	3	10	13	11	11.6	11.0	12.4	11.6	37.2	45.1
Juliana/Alba	3	10	13	9	10.1	9.0	11.7	10.3	25.8	37.9
Juliana/Alba	3	10	-7	11	5.9	10.9	-1.2	5.2	4.7	20.4
Juliana/Alba	3	10	-7	9	4.5	8.9	4.5	8.9	1.8	1.8
Juliana/Laura	3	2	13	4	6.7	4.3	10.1	6.9	9.3	37.2
Juliana/Laura	3	2	13	0	3.8	0.4	8.8	4.2	1.7	19.2
Juliana/Laura	3	2	-7	4	0.7	3.7	-3.4	0.4	4.0	21.9
Juliana/Laura	3	2	-7	0	-2.1	-0.2	-4.7	-2.3	30.5	39.0
Laura/Alba	2	10	4	11	4.7	10.8	6.3	8.7	3.9	9.9
Laura/Alba	2	10	4	9	4.5	8.9	5.6	7.4	3.7	10.0
Laura/Alba	2	10	0	11	1.0	10.7	3.6	7.4	0.8	4.5
Laura/Alba	2	10	0	9	0.8	8.8	2.9	6.1	1.4	8.1
Laura/Juliana	2	3	4	13	4.3	6.7	6.9	10.1	9.3	37.2
Laura/Juliana	2	3	4	-7	3.7	0.7	0.4	-3.4	4.0	21.9
Laura/Juliana	2	3	0	13	0.4	3.8	4.2	8.8	1.7	19.2
Laura/Juliana	2	3	0	-7	-0.2	-2.1	-2.3	-4.7	15.3	39.0
		Mean	5.0	5.0	5.8	5.8	5.0	5.0	9.1	23.7

Obs: y_{s_i} denotes the latent value in the i-th position in the sample

Note that the finite population mixed model predictor $\widehat{P}_i$ is unbiased but the standard linear mixed model $\widehat{Q}_i$ is not. We adopted the usual interpretation for $\widehat{Q}_i$, *i.e.*, as a predictor of the response for the i-th selected subject assuming that the associated variance corresponds to the subject-specific endogenous variance, which changes with the subject selected in the i-th position. However, we call the attention to the mistake in doing so, because according to the standard linear mixed model, the shrinkage constant $\gamma^2/(\gamma^2+\sigma_i^2)$ is attached to the position i in the sample and not to the subject selected in that position as in the example. This does not occur with the shrinkage constant $\gamma^2/(\gamma^2+\overline{\sigma}^2)$ considered in the finite population mixed model predictor. Nevertheless, the squared errors associated to the former are consistently smaller than the corresponding squared errors associated to the latter. The mean squared error of the finite population mixed model predictor is 23.7 while the mean squared error of the misinterpreted linear mixed model predictor is 9.1. This suggests that the unbiasedness condition considered in the derivation of $\widehat{P}_i$ may not be appropriate.

Extensive simulations were conducted by Moreno (2009) to examine the behaviour of both predictors under different setups involving underlying distributions as well as sample sizes. In general, the standard linear mixed model predictor performed better than the finite mixed model predictor.

4 Analysis of the cholesterol data in Table 1

In practical applications it is possible to fit finite population mixed models to data with endogenous and exogenous measurement errors using routines developed for standard mixed models and implemented in commonly used statistical software packages, as `SAS` or `R`.

The standard linear mixed model representation for the j-th measure of the i-th unit in the selected sample is

$$Y_{ij} = \mu + B_i + E_{ij}, \quad i = 1,\dots,n, \quad j = 1,\dots,n_i \tag{22}$$

with $B_i \overset{iid}{\sim} N(0,\gamma^2)$, and $E_{ij} \overset{iid}{\sim} N(0,\sigma_i^2)$ for heteroskedastic measurement errors or $E_{ij} \overset{iid}{\sim} N(0,\sigma^2)$ for homoskedastic measurement errors. The BLUP for $Y_i = \mu + B_i$ under this model has the form (19) in the homoskedastic case or (20) in the heteroskedastic case.

As an example of how the computation might be carried out, consider the data set described in the Introduction.

In Table 4 we display the the means of the 12 cholesterol measurements of each subject and assume, for illustrative purposes, that they are the corresponding "true" latent values. It follows that the "true" latent value variance is $\gamma^2 = (1/13)\sum_{s=1}^{13}(y_s - \overline{Y})^2 = 2939.9$ where $\overline{Y} = (1/13)\sum_{s=1}^{13} y_s$. Additionally, we let $\sigma_s^2 = (1/3)\sum_{q=1}^{4}(y_{sq} - y_s)^2$, $s = 1,\dots,13$ where y_{sq} denotes the mean cholesterol level of subject s in quarter q as the "true" variance of the endogenous measurement

Table 4 - Assumed latent value and endogenous measurement error variance along with predictors obtained under different variance structures

			`lme` predicted latent value		
Label	Latent	Endogenous	with error variance		
s	value (y_s)	variance (σ_s^2)	endogenous	exogenous	both
1	242.1	3545.1	241.0	241.5	238.7
2	263.2	1250.1	270.0	260.9	256.6
3	154.5	1932.9	157.9	158.3	164.8
4	232.6	3083.2	232.0	232.3	230.8
5	202.2	1299.4	203.2	203.3	205.1
6	280.4	1013.8	277.3	277.6	271.0
7	268.4	2216.9	265.9	267.1	261.0
8	198.4	1815.2	199.6	199.5	201.9
9	303.4	5222.3	299.1	300.9	290.4
10	237.7	548.9	236.8	237.2	235.0
11	141.8	1099.4	145.9	145.4	154.1
12	215.3	639.1	215.6	215.5	216.1
13	128.8	207.8	133.6	134.0	143.2

errors which are also presented in Table 4. Based on these values, the average "true" endogenous variance is $\overline{\sigma}^2 = (1/13)\sum_{s=1}^{13} \sigma_s^2 = 1836.5$.

In this setup, both the endogenous and exogenous measurement errors are heteroskedastic. Note that when only heteroskedastic endogenous measurement errors are present, the finite population mixed model predictor (19) has the same form as the standard linear mixed model predictor with homoscedastic measurement error variances.

Given that the columns labels in the data set are `Patient, Trim, Rep, Interv, Cholesterol`, the predictors may then be obtained by the `lme` function in `R` using the following commands:

```
require(nlme)
BLUP <- read.table("cholesterol.txt", header=T)
BLUP$Patient <- as.factor(BLUP$Patient)
BD1 <- groupedData(Cholesterol~1 | Patient, data = BLUP)
fit1<- lme(Cholesterol~1, data=BD1, random = ~1)
fit1$coefficients$fixed + fit1$coefficients$random$Patient
```

The `lme` predicted latent values are presented in the fourth column of Table 4. The restricted maximum likelihood estimated population variance (γ^2) is 2788.25. It underestimates the "true" variance (2939.9) by 5%. The `lme` estimated mean endogenous error variance ($\overline{\sigma}^2$) is 1836.47, and is practically equal to the "true" value (1836.5).

When only exogenous measurement errors are present, the finite population mixed model predictor (20) with heteroskedastic measurement error variance has the same form as the standard linear mixed model predictor with heteroskedastic measurement error variances. The corresponding predictors for the cholesterol example may be obtained via the following commands:

```
require(nlme)
BD1 <- groupedData(cholesterol~1 |Patient, data = dadoscol)
fit2 <- lme(Cholesterol~1, data=BD1, random = ~1,
          weights=varIdent(form = ~1|Interv))
fit2$coefficients$fixed + fit2$coefficients$random$Patient
```

The results are displayed in the fifth column of Table 4. The estimated population variance (γ^2) is 2797.15.

When both heterogeneous endogenous and exogenous measurement errors are present, the finite population mixed model predictors is equivalent to the standard linear mixed model predictors generated from the model

$$Y_{ijk} = \mu + B_i + D_{ik} + E_{ijk},\ i = 1, \ldots, n,\ k = 1, \ldots, n_i,\ j = 1, \ldots, p \qquad (23)$$

where $B_i \overset{iid}{\sim} N(0, \gamma^2)$, $D_{ik} \overset{iid}{\sim} N(0, \overline{\sigma}^2)$ and $E_{ijk} \overset{iid}{\sim} N(0, \sigma_j^2)$. In (23), D_{ik} represents the endogenous measurement error and E_{ijk} represents the exogenous measurement error.

We assume that different interviewers match the different evaluation conditions and consequently that the associated measurement errors may be considered as exogenous measurement errors. The corresponding "true" exogenous measurement error variances are presented in Table 5.

Table 5 - Exogenous measurement error variances

Position	Interviewer	Exogenous variance
1	CS	1012.8
2	KL	1623.3
3	SU	2001.1

The corresponding predictors for the cholesterol example may be obtained via the following commands:

```
BD4 <- groupedData(Cholesterol~Interv|Patient/Trim, data = BLUP)
fit3 <- lme(Cholesterol~1, data=BD4, random = ~1,
          weights=varIdent(form = ~1|Interv))
fit3$coefficients$fixed + fit3$coefficients$random$Patient
```

The `lme` predictors are displayed in the sixth column of Table 4.

The `lme` estimated population variance of the latent values is $\widehat{\gamma}^2 = 2455.1$; the `lme` estimate of the mean endogenous measurement error variance is $\widehat{\widehat{\sigma}} = 1312.8$ and the `lme` estimates of the exogenous measurement error variances are respectively, $\widehat{\widehat{\sigma}}_1^2 = 1017.8$, $\widehat{\widehat{\sigma}}_2^2 = 1592.8$ and $\widehat{\widehat{\sigma}}_3^2 = 2055.1$.

5 Discussion

By means of the example in Section 3, we showed that contrary to the usual interpretation, the heterogeneous standard linear mixed model predictor (21) does not take heterogeneous subject-specific (endogenous) variances into account. Since the step that links a unit label to its position in a response vector is omitted in the standard linear mixed model, this interpretation is erroneous. Finite population mixed models prevent such erroneous switch of concepts. This is aggravated by the fact that (21) corresponds to the BLUP obtained when exogenous heteroskedastic measurement errors are considered. By explicitly considering both types of measurement errors, we clarify this issue and extend the results of Singer *et al.* (2012).

Given that the expressions for best linear unbiased predictors for finite population mixed models may be matched to those obtained with standard linear mixed models with either homoskedastic, heteroskedastic (or both) measurement errors keeping the differences in interpretation in mind, we may use standard software designed for the latter to obtain predictors for the former. The advantage is that the covariance matrix is explicitly related to the exogenous or endogenous measurement errors so that the choice of the model may take advantage of the physical characteristics of the measurement process.

Finally, we note that neither model can account for the unit-specific endogenous measurement error variances when the interest is to predict the latent values of labelled selected units. One of the reasons for this may be related to the unbiasedness condition, which relates to overall expected response and not to the specific unit latent value. This issue has been raised by Robinson (1991) and by Buonaccorsi (2006) in a slightly different context.

An attempt to bypass this problem has been addressed by Stanek III and Singer (2004), who consider an expanded set of random variables following a random permutation probability distribution that keeps track of both the units labels and positions in the permutation. Unfortunately, under this model, the BLUP of a unit's parameter corresponds to the Horvitz-Thompson estimator when the unit is included in the sample, or zero otherwise and is subject to criticism.

Acknowledgments

This work was developed as part of the dissertation of my student and friend Germán Moreno. The paper was left unfinished because of his struggle to fight the cancer that so soon took him away from our company. I hope that my modest contribution will honour the memory of Germán.

We thank the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, grant # 304841/2019-6), Brazil for partial financial support.

We appreciate the careful revision from two anonymous referees.

MORENO, G.; SINGER, J. M.; STANEK III, E. J. Melhores preditores de valores latentes não tendenciosos lineares para modelos lineares de população finita com diferentes fontes de erro. *Rev. Bras. Biom.,* Lavras, v.39, n.4, p.571-586, 2021.

- *RESUMO: Desenvolvemos preditores lineares não enviesados ótimos (BLUP) para valores latentes de unidades amostrais rotuladas selecionadas de uma população finita na presença de duas fontes de erros de medida: endógenas, exógenas ou ambas. Parâmetros alvo usuais são a média populacional, o valor latente associado a uma unidade amostral rotulada ou o valor latente da unidade amostral selecionada numa determinada posição na amostra. Mostramos como os dois tipos de erros de medida afetam a matriz de covariâncias intraunidades amostrais e indicamos como o BLUP para populações finitas pode ser calculado por intermédio de* software *usualmente utilizado para ajustar modelos mistos com erros de medida endógenos ou exógenos, homocedásticos ou heterocedásticos.*

- *PALAVRAS-CHAVE: Erro de medição; matriz de covariância.*

References

BUONACCORSI, J. Estimation in two-stage models with heteroskedasticity. *International Statistical Review*, v.74, p.403-418, 2006.

COCHRAN, W. G. *Sampling techniques*, 3.ed., New York: Wiley, 1977. 472p.

DEMIDENKO, E. *Mixed models: theory and applications with R*, 2.ed., New York: Wiley, 2013. 768p.

DIGGLE, P.; HEARGERTY, P.; LIANG, K.; ZEGER, S. *Analysis of longitudinal data*, New York: Oxford University Press, 2002. 379p.

FITZMAURICE, G.; DAVIDIAN, M.; VERBEKE, G.; MOLENBERGHS, G. *Longitudinal data analysis: a handbook of modern statistical methods*, New York: Chapman & Hall, 2008. 617p.

MERRIAM, P. A.; OCKENE, I. S.; HEBERT, J. R.; ROSAL, M. C.; MATTHEWS, C. E. Seasonal variation of blood cholesterol levels: Study methodology. *Journal of Biologic Rhythms*, v.14, p.330-339, 1999.

MORENO, G. *Modelos mistos para populações finitas com erros de medida endógenos e exogenos*, 2009. 157p. Thesis (PhD). Departamento de Estatística, Universidade de São Paulo, São Paulo, 2009. URL: http://www.teses.usp.br/teses/disponiveis/45/45133/tde-29092009-195316/

ROBINSON, G. K. That blup is a good thing: the estimation of random effects. *Statistical Science*, v.6, p.15-51, 1991.

ROYALL, R. The linear least-squares prediction approach to two-stage sampling. *Journal of the American Statistical Association*, v.71, p.657-664, 1976.

SINGER, J. M.; STANEK III, E. J.; LENCINA, V. B.; GONZÁLEZ, L. M.; LI, W.; SAN MARTINO, S. Prediction with measurement error: do we really understand the blup? *Probability and Statistical Letters*, v.82, p.332-339, 2012.

STANEK III, E. J.; SINGER, J. M. Predicting random effects from finite population clustered samples with response error. *Journal of the American Statistical Association*, v.99, p.1119-1130, 2004.

SUKHATME, P. V. *Sampling theory of surveys applications*, 3.ed. Ames: Iowa State University Press. 1984. 532p.

SÄRNDAL, C. E.; SWENSSON, B.; WRETMAN, J. *Model assisted survey sampling*, New York: Springer-Verlag. 1992. 694p.

Received on 14.02.2021.

Approved after revised on 21.05.2021.