sciendo

ECONOMETRICS. EKONOMETRIA
Advances in Applied Data Analysis
Year 2020, Vol. 24, No. 2
ISSN 1507-3866; e-ISSN 2449-9994

ON EMPIRICAL BEST LINEAR UNBIASED PREDICTOR UNDER A LINEAR MIXED MODEL WITH CORRELATED RANDOM EFFECTS

Małgorzata K. Krzciuk
University of Economics in Katowice, Katowice, Poland
e-mail: malgorzata.krzciuk@uekat.pl
ORCID: 0000-0002-5906-5744

DOI: 10.15611/eada.2020.2.02
JEL Classification: C15, C51, C53

Abstract: The problem of small area prediction is considered under a Linear Mixed Model. The article presents a proposal of an empirical best linear unbiased predictor under a model with two correlated random effects. The main aim of the simulation analyses is a study of an influence of the occurrence of a correlation between random effects on properties of the predictor. In the article, an increase of the accuracy due to the correlation between random effects and an influence of model misspecification in cases of the lack of correlation between random effects are analyzed. The problem of the estimation of the Mean Squared Error of the proposed predictor is also considered. The Monte Carlo simulation analyses and the application were prepared in R language.

Keywords: Empirical Best Linear Unbiased Predictor, small area estimation, Monte Carlo simulation analyses.

1. Introduction

The problem of the prediction of small area total is considered. The main aim of the analyses in survey sampling was to propose an Empirical Best Linear Unbiased Predictor under the model with two correlated random effects, therefore the model approach problem was discussed.

In the simulation analyses the influence of the occurrence of a correlation between random effects on properties of the predictor was studies. The analyses concentrate on the increase of the accuracy due to the correlation between random effects and the influence of model misspecification in cases of the lack of correlation between random effects. In its application, the problem of the estimation of the Mean Squared Error of the proposed predictor was also considered.

In Sections 2 and 3 the General Linear Mixed Model with a special case regarding two correlated random effects is presented. Section 4 concerns the Empirical Best Linear Unbiased Predictor (EBLUP). In the next section, the problem of the estimation of the Mean Squared Error of the proposed predictor and its application are considered. Section 6 presents the results of the simulation study. In these analyses the proposal of EBLUP for the linear mixed model with two correlated random effects was considered. The last section is the conclusion of the application and simulation study.

2. General Linear Mixed Model

The analysed population Ω of size N is divided into D domains Ω_d, each of size N_d, where $d = 1,\dots,D$ and $N = \sum_{d=1}^{D} N_d$ ($\Omega = \bigcup_{d=1}^{D} \Omega_d$). Additionally, the sample s of size n and the sample in d-th domain of size n_d denoted by s_d are considered. The model which belongs to the class of the Linear Mixed Models (LMM) is analysed. The General Linear Mixed Model is given by:

$$\begin{cases} \mathbf{Y} = \mathbf{X\beta} + \mathbf{Zv} + \mathbf{e} \\ \mathrm{E}_{\xi}(\mathbf{v}) = \mathbf{0} \\ \mathrm{E}_{\xi}(\mathbf{e}) = \mathbf{0} \\ \mathrm{D}^2(\mathbf{v}) = \mathbf{G}(\boldsymbol{\delta}) \\ \mathrm{D}^2(\mathbf{e}) = \mathbf{R}(\boldsymbol{\delta}) \end{cases}, \tag{1}$$

where: $\mathbf{Y}$ – the random vector of values of the dependent variable, its distribution will be denoted by ξ, $\mathbf{X}$, $\mathbf{Z}$ – known matrices of auxiliary variables, $\boldsymbol{\beta}$ – the vector of unknown parameters.

Furthermore, $\mathbf{v}$ is a vector of random effects and $\mathbf{e}$ – vector of stochastic disturbances with variance-covariance matrices $\mathbf{G}$ and $\mathbf{R}$, respectively. Both matrices depend on a vector of unknown parameters $\boldsymbol{\delta}$ called variance components. The expected value relative to the ξ-distribution is denoted by $E_{\xi}(.)$ (cf. Jiang, 2007, pp. 1-2; Rao and Molina, 2015, p. 98). The variance-covariance matrix of $\mathbf{Y}$ under (1) is given by:

$$\mathbf{V}(\boldsymbol{\delta}) = \mathbf{ZG}(\boldsymbol{\delta})\mathbf{Z} + \mathbf{R}(\boldsymbol{\delta}). \tag{2}$$

The above model (1) can be also presented in the following form:

$$\mathbf{Y} = \mathbf{X\beta} + \mathbf{Z_1 v_1} + \mathbf{Z_2 v_2} + \cdots + \mathbf{Z_h v_h} + \mathbf{e}, \tag{3}$$

where:

$$\mathrm{D}^2\begin{bmatrix} \mathbf{v_1} \\ \mathbf{v_2} \\ \vdots \\ \mathbf{v_h} \end{bmatrix} = \begin{bmatrix} \mathbf{G_{11}} & \mathbf{G_{12}} & \cdots & \mathbf{G_{1h}} \\ \mathbf{G_{21}} & \mathbf{G_{22}} & \cdots & \mathbf{G_{2h}} \\ \cdots & \cdots & \cdots & \cdots \\ \mathbf{G_{h1}} & \cdots & \cdots & \mathbf{G_{hh}} \end{bmatrix}. \tag{4}$$

It should be noted that for $i \neq j$ it is possible that $\mathbf{G_{ij}} \neq \mathbf{0}$ and matrix $\mathbf{R}$ can be written as: $\mathbf{R}(\boldsymbol{\delta}) = \sigma_e^2 diag(v_i)$ for $1 \leq i \leq N$.

3. Some special cases of the Linear Mixed Model

In this section some special cases of the Linear Mixed Model are presented. The first group are models with one random effect. In this group two models can be analyzed: the nested error regression model and the model with a random slope. The first model was considered by Battese, Harter and Fuller (1988) and has the following form:

$$Y_{id} = \beta_1 x_{id} + \beta_0 + v_d + e_{id}. \tag{5}$$

Matrix **Z** for this model can be written as:

$$\mathbf{Z} = \begin{bmatrix} \mathbf{1_{N_1}} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1_{N_2}} & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \mathbf{0} & \cdots & \cdots & \mathbf{1_{N_D}} \end{bmatrix}_{\mathrm{N\times D}}. \tag{6}$$

Variance-covariance matrices for random effects and stochastic disturbance in this case have the following forms:

$$\mathbf{G}(\boldsymbol{\delta}) = \sigma_{v_d}^2 \mathbf{I}_{\mathrm{D\times D}}, \tag{7}$$

$$\mathbf{R}(\boldsymbol{\delta}) = \sigma_e^2 \mathrm{diag}(v_i) \text{ for } 1 \le \mathrm{i} \le \mathrm{N}, \tag{8}$$

so matrix **V** according to formula (2) can be written as:

$$\mathbf{V}(\boldsymbol{\delta}) = \mathrm{diag}_{1\le \mathrm{d} \le \mathrm{D}} \mathbf{V_d} = \mathrm{diag}_{1\le \mathrm{d} \le \mathrm{D}} \left(\sigma_{v_d}^2 \mathbf{1_{N_d}} \mathbf{1_{N_d}^T} + \sigma_e^2 \boldsymbol{I}_{\boldsymbol{N_d}\times \boldsymbol{N_d}}\right). \tag{9}$$

The same form of matrix **R** is assumed for all of the presented models.

The model with random slope analysed by Dempster, Rubin and Tsutakawa (1981) is given by:

$$Y_{id} = (\beta_1 + v_d) x_{id} + \beta_0 + e_{id}. \tag{10}$$

Matrix **Z** in this case has the following form:

$$\mathbf{Z} = \begin{bmatrix} \mathbf{x_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{x_2} & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \mathbf{0} & \cdots & \cdots & \mathbf{x_D} \end{bmatrix}_{\mathrm{N\times D}}, \tag{11}$$

where: $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_D}$ are the vectors of auxiliary variable for domains. Variance--covariance matrices **G** and **R** have the same form as for the first model but matrix **V** is given by:

$$\mathbf{V}(\boldsymbol{\delta}) = \mathrm{diag}_{1\le \mathrm{d} \le \mathrm{D}} \mathbf{V_d} = \mathrm{diag}_{1\le \mathrm{d} \le \mathrm{D}} \left(\sigma_{v_d}^2 \boldsymbol{x_d} \boldsymbol{x_d^T} + \sigma_e^2 \boldsymbol{I}_{\boldsymbol{N_d}\times \boldsymbol{N_d}}\right). \tag{12}$$

Furthermore, models with two random effects – uncorrelated or correlated – are discussed. It is assumed that both of the random effects are specific for domains. The model with two uncorrelated random effects has the following form:

$$Y_{id} = (\beta_1 + v_{2d})x_{id} + \beta_0 + v_{1d} + e_{id}. \quad (13)$$

The matrix of auxiliary variables **Z** in this case has a more complex form than for the model with only one random effect:

$$\mathbf{Z} = \begin{bmatrix} \mathbf{1}_{N_1} & \mathbf{x_1} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{1}_{N_2} & \mathbf{x_2} & \cdots & \mathbf{0} & \mathbf{0} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}_{N_D} & \mathbf{x_D} \end{bmatrix}_{N \times 2D}. \quad (14)$$

The variance-covariance matrix of random effects in this case is given by:

$$\mathbf{G}(\boldsymbol{\delta}) = \begin{bmatrix} \mathbf{G_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{G_2} & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \mathbf{0} & \cdots & \cdots & \mathbf{G_D} \end{bmatrix}_{2D \times 2D}, \quad (15)$$

where each of blocks has the following form: $\mathbf{G_D} = \begin{bmatrix} \sigma^2_{v_{1d}} & 0 \\ 0 & \sigma^2_{v_{2d}} \end{bmatrix}$. By inserting this matrices into formula (2), the following covariance matrix of **Y** is obtained:

$$\mathbf{V}(\boldsymbol{\delta}) = diag_{1 \le d \le D}\mathbf{V_d} =$$
$$= diag_{1 \le d \le D}(\sigma^2_{v_d}\mathbf{1}_{N_d}\mathbf{1}^{\boldsymbol{T}}_{n_d} + \sigma^2_{v_{2d}}\mathbf{x}_d\mathbf{x}^{\boldsymbol{T}}_d + \sigma^2_e\mathbf{I}_{N_d \times N_d}). \quad (16)$$

It should be noted that in this case matrix **V** is the sum of matrices (9) and (12). If correlation between random effects is assumed, the model is given by:

$$Y_{id} = (\beta_1 + v^*_{2d})x_{id} + \beta_0 + v^*_{1d} + e_{id}. \quad (17)$$

The variance-covariance matrix, similarly to the previous model, is block-diagonal but the blocks have the following form:

$$\mathbf{G_d} = \begin{bmatrix} \sigma^2_{v^*_{1d}} & \rho\sigma_{v^*_{1d}}\sigma_{v^*_{2d}} \\ \rho\sigma_{v^*_{1d}}\sigma_{v^*_{2d}} & \sigma^2_{v^*_{2d}} \end{bmatrix}. \quad (18)$$

The variance-covariance matrix of **V** has a fairly similar form compared to model (13) but an additional, third element with parameter ρ was observed:

$$\mathbf{V}(\boldsymbol{\delta}) = diag_{1 \le d \le D}\mathbf{V_d} = diag_{1 \le d \le D}(\sigma^2_{v_d}\mathbf{1}_{N_d}\mathbf{1}^{\boldsymbol{T}}_{N_d} + \sigma^2_{v_{2d}}\mathbf{x}_d\mathbf{x}^{\boldsymbol{T}}_d +$$
$$+\rho\sigma_{v^*_{1d}}\sigma_{v^*_{2d}}(\mathbf{1_{N_d}x^T_d}+\mathbf{x_d1^T_{N_d}}) + \sigma^2_e\mathbf{I_{N_d \times N_d}}). \quad (19)$$

Figure 1 and Figure 2 present slopes and intercepts for models with two uncorrelated and correlated random effects, respectively. Each line corresponds to one domain.

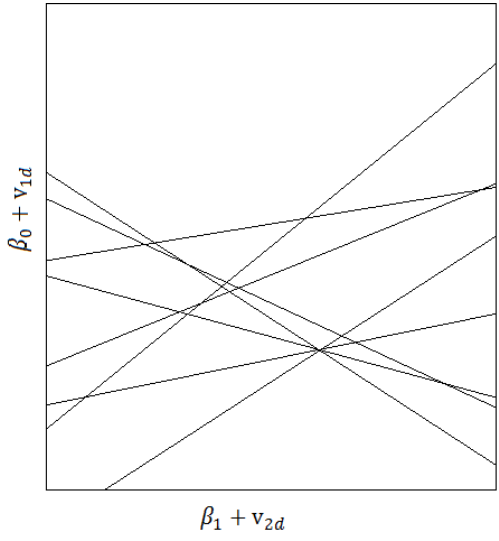

Fig. 1. Slopes and intercepts for a model with uncorrelated random effects

Source: own elaboration.

A clear difference can be seen in the line layout in these two cases. In Figure 2 the signs of the slopes for all the lines are the same, and most of the lines intersect at one point.

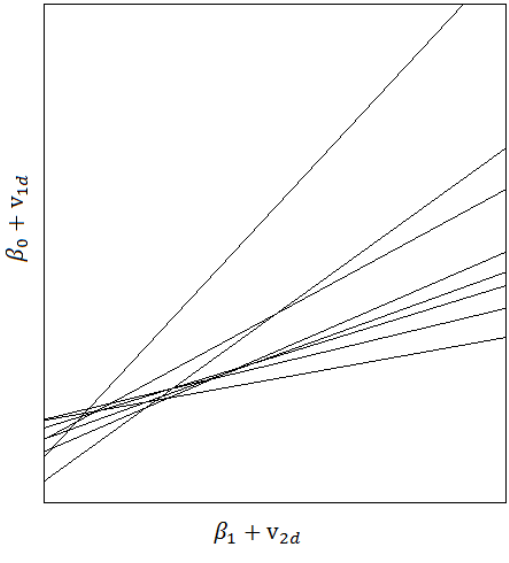

Fig. 2. Slopes and intercepts for a model with correlated random effects

Source: own elaboration.

Figure 3 and Figure 4 show a graphic presentation of matrix G for the considered two models.

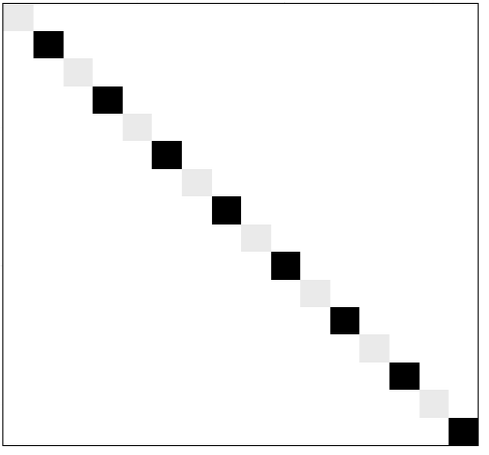

Fig. 3. Matrix G for a model with two uncorrelated random effects

Source: own elaboration.

On these figures each of the squares represents the value of one element in the matrix. Non-zero matrix elements are marked in grey. The darker colour of the square means a higher value. In Figure 4 it should be noted that covariance within domains is non-zero.

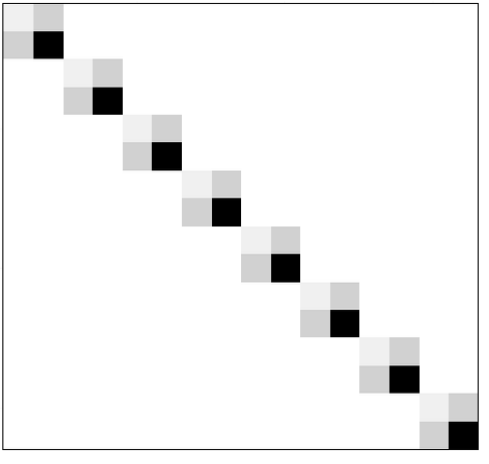

Fig. 4. Matrix G for a model with two correlated random effects

Source: own elaboration.

Obviously, more complex models can be also considered – LMMs with three or more random effects and with more than two correlated random effects.

It should be noted that LMMs with correlated random effects found their application in e.g.: the estimation of plasma concentration of a drug by a nonlinear mixed effects model (Dumont, Chenel, and Mentre, 2014), analyses of health care costs at the end of life (Menec et al., 2004) and analyses of the bias and the precision of the estimates for pharmacokinetics (PK) and pharmacodynamics (PD) (Ogungbenro, et al., 2008). These models may find their application also in other areas, including economics.

4. Empirical Best Linear Unbiased Predictor

The problem of the prediction of some characteristic $\theta = \gamma^T \mathbf{Y}$ was also considered. Additionally the following decomposition of vector Y was assumed:

$$\mathbf{Y} = [\mathbf{Y_s^T} \quad \mathbf{Y_r^T}]^\mathbf{T}, \tag{20}$$

where $\mathbf{Y_s^T}$ is the vector of size n, for elements which were drawn to sample s, the vector of size $\mathrm{N}_r = N - n$, $\mathbf{Y_r^T}$ corresponds to elements not drawn to the sample. Similarly, vector $\boldsymbol{\gamma}$ can be decomposed as follows:

$$\boldsymbol{\gamma} = [\boldsymbol{\gamma_s^T} \quad \boldsymbol{\gamma_r^T}]^\mathbf{T}, \tag{21}$$

where for the total value in the d-th domain the k-th element of the vector $\boldsymbol{\gamma}$ equals 1 for $k \in \Omega_d$ and 0 otherwise. The variance-covariance matrix of $\mathbf{Y}$ is given by:

$$\mathbf{V}(\boldsymbol{\delta}) = \mathrm{D}^2(\mathbf{Y}) = \mathrm{D}^2\begin{bmatrix}\mathbf{Y_s}\\ \mathbf{Y_r}\end{bmatrix} = \begin{bmatrix}\mathbf{V_{ss}}(\boldsymbol{\delta}) & \mathbf{V_{sr}}(\boldsymbol{\delta})\\ \mathbf{V_{rs}}(\boldsymbol{\delta}) & \mathbf{V_{rr}}(\boldsymbol{\delta})\end{bmatrix}. \tag{22}$$

According to the Royall (1976) theorem, the Best Linear Unbiased Predictor is given by:

$$\hat{\theta}_{\mathrm{BLUP}} = \boldsymbol{\gamma_s^T}\mathbf{Y_s} + \boldsymbol{\gamma_r^T}\left[\mathbf{X_r}\widehat{\boldsymbol{\beta}}(\boldsymbol{\delta}) + \mathbf{V_{rs}}(\boldsymbol{\delta})\mathbf{V_{ss}^{-1}}(\boldsymbol{\delta})\left(\mathbf{Y_s} - \mathbf{X_s}\widehat{\boldsymbol{\beta}}(\boldsymbol{\delta})\right)\right]. \tag{23}$$

If the diagonal form of matrix $\mathbf{R}(\boldsymbol{\delta})$ is assumed, predictor (23) simplifies to (cf. Żądło, 2017):

$$\hat{\theta}_{\mathrm{BLUP}} = \boldsymbol{\gamma_s^T}\mathbf{Y_s} + \boldsymbol{\gamma_r^T}\mathbf{X_r}\widehat{\boldsymbol{\beta}}(\boldsymbol{\delta}) + \boldsymbol{\gamma_r^T}\mathbf{Z_r}\hat{\mathbf{v}}, \tag{24}$$

where $\widehat{\boldsymbol{\beta}}$ and $\hat{\mathbf{v}}$ are vectors of estimates of fixed and random effects, respectively. Additionally, if $\boldsymbol{\delta}$ is replaced by its estimator, a two stage predictor called EBLUP was obtained. In the Monte Carlo simulation analyses and application of the proposal of EBLUP for model (17) were considered.

5. The MSE of the EBLUP and its estimators – application

In the following application the problem of estimation of the MSE of the analysed predictor was considered. The prediction Mean Squared Error (MSE) of the BLUP was given by (cf. Royall, 1976):

$$MSE_{\xi}\left(\hat{\theta}_{BLUP}\right) = g_1(\boldsymbol{\delta}) + g_2(\boldsymbol{\delta}), \tag{25}$$

where:

$$g_1(\boldsymbol{\delta}) = \gamma_r^T\left(\boldsymbol{V}_{rr}(\boldsymbol{\delta}) - \boldsymbol{V}_{rs}(\boldsymbol{\delta})\boldsymbol{V}_{ss}^{-1}(\boldsymbol{\delta})\boldsymbol{V}_{sr}(\boldsymbol{\delta})\right)\gamma_r \tag{26}$$

and

$$g_2(\boldsymbol{\delta}) = \gamma_r^T(\mathbf{X}_r - \boldsymbol{V}_{rs}(\boldsymbol{\delta})\boldsymbol{V}_{ss}^{-1}(\boldsymbol{\delta})\mathbf{X}_s)(\mathbf{X}_s^T\boldsymbol{V}_{ss}^{-1}(\boldsymbol{\delta})\mathbf{X}_s)^{-1} \times \\ \times (\mathbf{X}_r - \boldsymbol{V}_{rs}(\boldsymbol{\delta})\boldsymbol{V}_{ss}^{-1}(\boldsymbol{\delta})\mathbf{X}_s)^T\gamma_r. \tag{27}$$

The MSE of the EBLUP has the following form (Datta and Lahiri, 2000):

$$MSE_{\xi}\left(\hat{\theta}_{EBLUP}\right) = g_1(\boldsymbol{\delta}) + g_2(\boldsymbol{\delta}) + g_3(\boldsymbol{\delta}) + o(D^{-1}), \tag{28}$$

where $g_1(\boldsymbol{\delta})$ and $g_2(\boldsymbol{\delta})$ are given by (26) and (27), respectively, and the last element is given by:

$$g_3(\boldsymbol{\delta}) = tr\left(\frac{\partial \boldsymbol{c}^T}{\partial \boldsymbol{\delta}}\boldsymbol{V}_{ss}(\boldsymbol{\delta})\frac{\partial \boldsymbol{c}^T}{\partial \boldsymbol{\delta}}\breve{D}^2\left(\widehat{\boldsymbol{\delta}}\right)\right), \tag{29}$$

where: $\boldsymbol{c}^T = \gamma_r^T\boldsymbol{V}_{rs}(\boldsymbol{\delta})\boldsymbol{V}_{ss}^{-1}(\boldsymbol{\delta})$ and $\breve{D}^2\left(\widehat{\boldsymbol{\delta}}\right)$ is the asymptotic variance-covariance matrix of estimator $\widehat{\boldsymbol{\delta}}$.

In the application, three estimators of (28) were considered. The first of them is a classic estimator called naive given by (Kackar and Harville 1984, pp. 854-855):

$$\hat{MSE}_{\xi N}\left(\hat{\theta}_{EBLUP}\right) = g_1\left(\widehat{\boldsymbol{\delta}}\right) + g_2\left(\widehat{\boldsymbol{\delta}}\right). \tag{30}$$

where $g_1\left(\widehat{\boldsymbol{\delta}}\right)$ and $g_2\left(\widehat{\boldsymbol{\delta}}\right)$ can be calculated using formulas (26) and (27) where $\boldsymbol{\delta}$ is replaced by its estimate. Additionally, for this estimator:

$$E_{\xi}\left(\hat{MSE}_{\xi N}\left(\hat{\theta}_{EBLUP}\left(\widehat{\boldsymbol{\delta}}\right)\right)\right) - MSE_{\xi}\left(\hat{\theta}_{EBLUP}\left(\widehat{\boldsymbol{\delta}}\right)\right) = O(D^{-1}). \tag{31}$$

where $MSE_{\xi}(.)$ is ξ-Mean Squared Error of the predictor.

The next two MSE estimators were based on the parametric bootstrap method. These estimators are based on the following bootstrap model (cf. Chatterjee, Lahiri, Li, 2008, pp. 1229-1230):

$$\mathbf{Y}^* = \mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{Z}\mathbf{v}^* + \mathbf{e}^*, \tag{32}$$

where:

- $\widehat{\boldsymbol{\beta}}$ is the LS estimator of $\boldsymbol{\beta}$,
- $\mathbf{v}^* \sim \mathbf{N}\left(\mathbf{0}, \mathbf{G}\left(\widehat{\boldsymbol{\delta}}\right)\right)$ and $\mathbf{e}^* \sim \mathbf{N}\left(\mathbf{0}, \mathbf{R}\left(\widehat{\boldsymbol{\delta}}\right)\right)$,
- $\widehat{\boldsymbol{\delta}}$ is the REML or ML estimator of $\boldsymbol{\delta}$.

The parametric bootstrap estimator considered by Gonzales-Manteiga (2008) is given by:

$$\widehat{MSE}_{\xi boot}\left(\hat{\theta}_{EBLUP}\right) = E_*\left(\hat{\theta}_{EBLUP}\left(\widehat{\boldsymbol{\beta}}\left(\widehat{\boldsymbol{\delta}}^*\right), \widehat{\boldsymbol{\delta}}^*\right) - \theta^*\right)^2 =$$
$$= B^{-1}\sum_{b=1}^{B}\left(\hat{\theta}_{EBLUP}\left(\widehat{\boldsymbol{\beta}}\left(\widehat{\boldsymbol{\delta}}^{*(b)}\right), \widehat{\boldsymbol{\delta}}^{*(b)}\right) - \theta^{*(b)}\right)^2, \tag{33}$$

where:

- $\widehat{\boldsymbol{\delta}}$ and $\widehat{\boldsymbol{\beta}}$ are REML estimators;
- $\widehat{\boldsymbol{\delta}}^{*(b)}$ is given by the same formula as $\boldsymbol{\delta}$ where $\mathbf{Y}$ is replaced by $\mathbf{Y}^*$.

Additionally, $E_*(.)$ is the expected value in the bootstrap distribution and $\theta^{*(b)}$ is the value of θ obtained in the *b*-th realization of the bootstrap model.

The last is the estimator proposed by Butar and Lahiri (2003), which has the following form:

$$\widehat{MSE}_{\xi boot-BL}\left(\hat{\theta}_{EBLUP}\right) = g_1\left(\widehat{\boldsymbol{\delta}}\right) + g_2\left(\widehat{\boldsymbol{\delta}}\right) +$$
$$+E_*\left(g_1\left(\widehat{\boldsymbol{\delta}}^*\right) + g_2\left(\widehat{\boldsymbol{\delta}}^*\right) - \left(g_1\left(\widehat{\boldsymbol{\delta}}\right) + g_2\left(\widehat{\boldsymbol{\delta}}\right)\right)\right) + \tag{34}$$
$$+E_*\left(\hat{\theta}_{EBLUP}\left(\widehat{\boldsymbol{\beta}}\left(\widehat{\boldsymbol{\delta}}^*\right), \widehat{\boldsymbol{\delta}}^*\right) - \theta^*\right)^2,$$

where $g_1\left(\widehat{\boldsymbol{\delta}}^*\right)$ and $g_2\left(\widehat{\boldsymbol{\delta}}^*\right)$ are calculated based on (26) and (27) where $\boldsymbol{\delta}$ is replaced by $\widehat{\boldsymbol{\delta}}^*$. This estimator is asymptotically unbiased in the following sense:

$$E_\xi\left(\widehat{MSE}_{\xi N}\left(\hat{\theta}_{EBLUP}\left(\widehat{\boldsymbol{\delta}}\right)\right)\right) - MSE_\xi\left(\hat{\theta}_{EBLUP}\left(\widehat{\boldsymbol{\delta}}\right)\right) = o(D^{-1}). \tag{35}$$

In the application dataset from Särndal, Swensson and Wretman (1992), concerning Swedish municipalities, ($N = 284$) was used. The dependent variable in the considered model (17) were revenues from 1985 municipal taxation (in millions of kronor), and as the auxiliary variable population in 1975 (in thousands of people) was used. The division of population into eight domains, according to the region was considered. The sample size $n = 28$ ($\sim 10\%$ of the population size) was drawn using the Brewer sampling scheme. The sample size results from the number of elements in the domains. The considered characteristic is the total value in a domain. For parametric bootstrap estimators, $B = 200$ was assumed. In Table 1 the results of the application are presented – values of the relative Root Mean Squared Error ($\widehat{rRMSE}$) for each of the eight domains.

For each of the estimators, the value of $\widehat{rRMSE}$ is not higher than 22%. It should be noted that the naïve estimator has a higher order of the bias than the estimator

Table 1. Results of the application for 8 domains – values of $\hat{rRMSE}$ (in %)

Domain	1	2	3	4	5	6	7	8
$\hat{MSE}_{\xi N}(\hat{\theta}_{EBLUP})$	4.61	6.97	5.76	7.16	6.97	8.92	8.48	14.66
$\hat{MSE}_{\xi boot}(\hat{\theta}_{EBLUP})$	6.09	3.07	3.87	4.12	3.82	8.07	5.37	15.13
$\hat{MSE}_{\xi boot-BL}(\hat{\theta}_{EBLUP})$	6.77	5.80	10.35	5.64	5.74	10.34	17.22	21.08

Source: own elaboration.

considered by Butar and Lahiri (2003). Furthermore, the order of the bias for the estimator presented by Gonzales-Manteiga, et. al. (2008) is unknown. Taking into account the properties of the considered estimators, it is recommended to use estimator $\hat{MSE}_{\xi boot-BL}$.

6. Simulation study

In the simulation study the same dataset and the same problem as in application were considered. The choice was made from five models:

- linear regression model with one dependent variable and intercept:

$$Y_{id} = \beta_1 x_{id} + \beta_0 + e_{id}, \tag{36}$$

- nested error regression model (5),
- model with random slope (10),
- linear mixed model with two uncorrelated random effects (13),
- linear mixed model with two correlated random effects (17).

Based on the Akaike Information Criterion (AIC) (see Biecek, 2012, p. 123), the last model was chosen – the Linear Mixed Model with two correlated random effects specific for domains. To verify the model significance tests of fixed effects and variance of random components can be used for fixed effects – a permutation version of the conditional t test (Wolfinger, 1993), or a permutation version of the test based on the likelihood function (Biecek, 2012); and for random components, among others, a permutation version of the test based on the likelihood function (Biecek, 2012). The properties of these tests were considered in simulation studies e.g. by Krzciuk and Żądło (2014a, b).

The simulation study can be divided into two parts. In both of them a model-based approach was assumed. In the first part data were generated based on the model with correlated random effects, in the second – those with uncorrelated random effects. In both parts two predictors were considered:

- EBLUP_1 – the proposed predictor based on the model with correlated random effects, where model parameters are replaced by their estimates;

- EBLUP_2 – the predictor based on the model with uncorrelated random effects, where model parameters are replaced by their estimates.

The number of Monte Carlo iterations is 2000.

In Table 2 results of the first part of the simulation study are presented – the values of relative biases of the considered EBLUPs and the ratios of their MSEs. In this case generated data are based on the model with correlated random effects. Each column of the table shows the results for one of the eight domains.

Table 2. Results of the 1st part of simulation study – model-based approach ($\rho \neq 0$)

Domain	1	2	3	4	5	6	7	8
$\text{rB}(\text{EBLUP}_1)$ (in %)	–0.01	–0.14	0.38	0.05	–0.13	0.05	–0.15	–0.65
$\text{rB}(\text{EBLUP}_2)$ (in %)	–0.07	–0.27	0.37	0.05	–0.08	0.12	–0.53	–0.12
$\frac{\text{MSE}(\text{EBLUP}_1)}{\text{MSE}(\text{EBLUP}_2)}$	0.61	0.73	0.93	0.80	0.73	0.96	0.72	0.83

Source: own elaboration.

In most cases the absolute values of relative bias ($rB(.)$) for the proposed EBLUP were lower or quite similar to EBLUP_2. The absolute values of the relative bias for both of the predictors were not higher than 1%. The increase of accuracy of the proposed EBLUP compared to EBLUP_2, due to the correlation between random effects was between 4% and 39%.

Table 3 presents the results of the second part of the study, where data were generated based on the model with uncorrelated random effects.

Table 3. Results of the 2nd part of the simulation study – model-based approach ($\rho = 0$)

Domain	1	2	3	4	5	6	7	8
$\text{rB}(\text{EBLUP}_1)$ (in %)	0.67	0.29	–0.57	–0.58	0.09	–1.22	0.39	0.74
$\text{rB}(\text{EBLUP}_2)$ (in %)	–0.67	0.39	–0.52	–0.61	0.37	–1.14	–0.24	–1.01
$\frac{\text{MSE}(\text{EBLUP}_1)}{\text{MSE}(\text{EBLUP}_2)}$	0.79	0.83	1.05	0.99	1.20	1.00	0.98	0.96

Source: own elaboration.

In this part of the analyses $rB(.)$ for EBLUP for models with correlation are quite similar to EBLUP_2. The loss of accuracy resulting from the model misspecification for the proposed EBLUP was not higher than 5%, except for only one case – the fifth domain.

7. Conclusion

In the article, the EBLUP for the linear mixed model with two correlated random effects was proposed. In the application, the problem of the estimation of the MSE of the analysed predictor was considered. In the analyses, the classic naive estimator and two estimators based on parametric bootstrap method were taken into account. The obtained results suggest using estimator $\hat{MSE}_{\xi boot-BL}$, but this problem requires further research.

The main aim of the simulation study was a comparison of the properties of the proposed predictor and EBLUP based on the model with uncorrelated random effects. In most cases, for the considered dataset the absolute values of the relative bias for the proposed EBLUP were lower or quite similar to EBLUP based on the model with uncorrelated random effects. The increase of accuracy for the proposed EBLUP due to correlation between random effects, in most cases was higher than 10%. The loss of accuracy resulting from model misspecification was not higher than 5%, except for only one domain. Even if the lack of the correlation between random effects is assumed, the EBLUP under the model, where the correlation is taken into account, has good properties.

This paper was presented at the conference MSA 2019 which financed its publication. The organization of the international conference "Multivariate Statistical Analysis 2019" (MSA 2019) was supported from resources for the popularization of scientific activities from the Minister of Science and Higher Education in the framework of agreement No 712/P-DUN/202019.

References

Battese, G. E., Harter, R. M., and Fuller, W. A. (1988). An error-components model for prediction of count crop area using survey satellite data. *Journal of the American Statistical Association*, *83*(401), 28-36.

Biecek, P. (2012). *Analiza danych z programem R. Modele liniowe z efektami stałymi i losowymi i mieszanymi*. Warszawa: Wydawnictwo Naukowe PWN.

Butar, F. B., and Lahiri, P. (2003). On measures of uncertainty of empirical Bayes small-area estimators. *Journal of Statistical Planning and Inference*, *112*(1-2), 635-676.

Chatterjee, S., Lahiri, P., and Li, H. (2008). Parametric bootstrap approximation to the distribution of EBLUP and related prediction intervals in linear mixed models. *The Annals of Statistics*, *36*(2), 1221-1245.

Datta, G. S., and Lahiri, P. (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statistica Sinica*, (10), 613-627.

Dempster, A. P., Rubin, D. B., and Tsutakawa, R. K. (1981). Estimation in covariance components models. *Journal of the American Statistical Association*, *76*(374), 341-353.

Dumont, C., Chenel, M., and Mentre, F. (2014). Influence of covariance between random effects in design for nonlinear mixed-effect models with an illustration in pediatric pharmacokinetics. *Journal of Biopharmaceutical Statistics*, *24*(3), 471-492.

Gonzales-Manteiga, W., et. al. (2008). Bootstrap mean squared error of a small-area EBLUP. *Journal of Statistical Computation and Simulation*, (78), 443-462.

Jiang, J. (2007), *Linear and Generalized Linear Mixed Models and their applications.* New York: Springer Science+Business Media.

Kackar, R. N., and Harville, D. A. (1994). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistical Association*, (79), 853-862.

Krzciuk, M., and Żądło, T. (2014a). On some tests of variance components for linear mixed models. *Studia Ekonomiczne – Zeszyty Naukowe Uniwersytetu Ekonomicznego w Katowicach*, (189), 77-85.

Krzciuk, M., and Żądło, T. (2014b). On some tests of fixed effects for linear mixed models. *Studia Ekonomiczne – Zeszyty Naukowe Uniwersytetu Ekonomicznego w Katowicach*, (189), 49-57.

Menec, V., et. al. (2004). *Patterns of health care use and cost at the end of life*. Winnipeg: MB: Manitoba Centre for Health Policy.

Ogungbenro, K., et. al. (2008). Incorporating correlation in interindividual variability for the optimal design of multiresponse pharmacokinetic experiments. *Journal of Biopharmaceutical Statistics*, *18*(2), 342-358.

Rao, J. N. K., and Molina, I. (2015). *Small area estimation.* Hoboken, New Jersey: John Wiley and Sons.

Royall, R. M. (1976). The linear least-squares prediction approach to two-stage sampling. *Journal of the American Statistical Association*, *71*(355), 657-664.

Särndal, C. E., Swensson, B., and Wretman, J. (1992). *Model assisted survey sampling.* New York: Springer Verlag.

Wolfinger, R. (1993). Covariance structure selection in general mixed models. *Communications in Statistics – Simulation and Computation*, *22*(4), 1079-1106.

Żądło, T. (2017), On prediction of population and subpopulation characteristics for future periods. *Communications in Statistics – Simulation and Computation*, *46*(10), 8086-8104.

O EMPIRYCZNYM NAJLEPSZYM LINIOWYM NIEOBCIĄŻONYM PREDYKTORZE DLA PEWNEGO MODELU MIESZANEGO

Streszczenie: Zagadnieniem poruszanym w artykule jest problem predykcji w przypadku pewnego modelu należącego do klasy liniowych modeli mieszanych. W opracowaniu została przedstawiona propozycja empirycznego najlepszego liniowego nieobciążonego predyktora dla liniowego modelu mieszanego z dwoma skorelowanymi efektami losowymi. Głównym celem opracowania jest symulacyjne zbadanie wpływu występowania zależności między efektami losowymi na własności rozważanego predyktora. W artykule podjęto również problem estymacji błędu średniokwadratowego zaproponowanego predyktora. Badanie symulacyjne oraz przykład przygotowano z użyciem programu R.

Słowa kluczowe: empiryczny najlepszy liniowy nieobciążony predyktor, statystyka małych obszarów, badanie symulacyjne.

Quote as: Krzciuk, M. K. (2020). On empirical best linear unbiased predictor under A linear mixed model with correlated random effects. *Econometrics. Ekonometria. Advances in Applied Data Analysis*, *24*(2).