ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 269, 2012

Tomasz Żądło[*]

ON PSEUDO-EBLUP UNDER SOME MODEL FOR LONGITUDINAL DATA WITH AUXILIARY VARIABLES

Abstract. The problem of modeling longitudinal profiles is considered assuming that the population and elements affiliation to subpopulations may change in time. The considerations are based on a model with auxiliary variables for longitudinal data with element and subpopulation specific random components (compare Verbeke, Molenberghs, 2000; Hedeker, Gibbons, 2006) which is a special case of the General Linear Model (GLM) the General Linear Mixed Model (GLMM). In the paper the pseudo-empirical best linear unbiased predictor (Pseudo-EBLUP) based on model-assisted approach will be presented along with its mean squared error (MSE) and its estimators. In the simulation study its accuracy will be compared with some calibration estimators which are based on model-assisted approach too.

Key words: small area estimation, pseudo-empirical best linear unbiased predictors, longitudinal data

I. INTRODUCTION

In the paper longitudinal data for periods t=1,...,M are considered. In the period t the population of size N_t is denoted by Ω_t. The population in the period t is divided into D disjoint domains (subpopulations) Ω_{dt} of size N_{dt}, where d=1,...,D. Let the set of population elements for which observations are available in the period t be denoted by s_t and its size by n_t. The set of domain elements for which observations are available in the period t is denoted by s_{dt} and its size by n_{dt}.

We assume that the population may change in time and that one population element may change its domain affiliation in time (from technical point of view observations of some population element which change its domain affiliation are treated as observations of new population element). It means that i and t completely identify domain affiliation but additional subscript d will be needed as well. Let M_{id} denote the number of periods when the i-th population element

[*] Ph.D., Department of Statistics, University of Economics in Katowice.

belongs to the d-th domain. Let us denote the number of periods when the i-th population element (which belongs to the d-th domain) is observed by m_{id}.

Values of the variable of interest are realizations of random variables Y_{idj} for the i-th population element which belongs to the d-th domain in the period t_{ij}, where i=1,...,N, j=1,...,M_{id}, d=1,...,D. The vector of size $M_{id}\times 1$ of random variables Y_{idj} for the i-th population element from the d-th domain will be denoted by $\mathbf{Y_{id}}=\left[Y_{idj}\right]$, where $j=1,...,M_{id}$.

We consider two-stage superpopulation model used for longitudinal data (compare Verbeke, Molenberghs (2000); Hedeker, Gibbons (2006)). Firstly:

$$\mathbf{Y_{id}}=\mathbf{Z_{id}}\boldsymbol{\beta}_{\mathbf{id}}+\mathbf{e_{id}}, \tag{1}$$

where i=1,...,N; d=1,...,D, $\mathbf{Z_{id}}$ is known matrix of size $M_{id}\times q$, $\boldsymbol{\beta}_{\mathbf{id}}$ is a vector of unknown parameters of size $q\times 1$, $\mathbf{e_{id}}$ is a random component vector of size $M_{id}\times 1$. Vectors $\mathbf{e_{id}}$ (i=1,...,N; d=1,...,D) are independent with **0** vector of expected values and variance-covariance matrix $\mathbf{R_{id}}$. Although $\mathbf{R_{id}}$ may depend on i it is often assumed that $\mathbf{R_{id}}=\sigma_e^2\mathbf{I_{M_{id}}}$ where $\mathbf{I_{M_{id}}}$ is the identity matrix of rank M_{id}. Secondly, we assume that:

$$\boldsymbol{\beta}_{\mathbf{id}}=\mathbf{K_{id}}\boldsymbol{\beta}+\mathbf{v_{id}}, \tag{2}$$

where i=1,...,N; d=1,...,D, $\mathbf{K_{id}}$ is known matrix of size $q\times p$, $\boldsymbol{\beta}$ is a vector of unknown parameters of size $p\times 1$, $\mathbf{v_{id}}$ is a vector of random components of size $q\times 1$. It is assumed that vectors $\mathbf{v_{id}}$ (i=1,...,N; d=1,...,D) are independent with **0** vector of expected values and variance-covariance matrix $\mathbf{G_{id}}=\mathbf{H}$ what means that $\mathbf{G_{id}}$ does not depend on i.

Similar assumptions to (1) and (2) are presented by Verbeke, Molenberghs (2000) p. 20 but there are 3 differences. Firstly, in the book assumptions are made for profiles defined by elements. In this paper assumptions are made for profiles defined by elements and domain affiliation i.e. $\mathbf{Y_{id}}$ (of size $M_{id}\times 1$). Secondly, in the book the assumptions are made only for the sampled elements. In this paper they are made for all of population elements. Thirdly, Verbeke and

Molenberghs (2000) (unlike in this paper) in their notations do not take the possibility of population changes in time into account.

Based on (1) and (2) it is obtained that:

$$\mathbf{Y_{id}} = \mathbf{X_{id}}\boldsymbol{\beta} + \mathbf{Z_{id}}\mathbf{v_{id}} + \mathbf{e_{id}}, \tag{3}$$

where i=1,...,N; d=1,...,D, $\mathbf{X_{id}} = \mathbf{Z_{id}}\mathbf{K_{id}}$ is known matrix of size $M_{id} \times p$.

Let us consider two special cases of the model (3). The model which will be called special case 1 is a random regression coefficient model similar to the one proposed by Depmster, Rubin and Tsutakawa (1981) and studied later e.g. by Prasad and Rao (1990) or Moura and Holt (1999) (but they do not consider longitudinal model, and they consider domain specific v random component). We assume that:

$$Y_{idj} = (\beta_d + v_{id})x_{ij} + e_{idj} = \beta_d x_{ij} + v_{id}x_{ij} + e_{idj} \tag{4}$$

where i=1,...,N; d=1,...,D, j=1,...,M_{id}. In the considered model we assume that (compare Verbeke and Molenberghs, 2000) $\mathbf{R_{id}} = \sigma_e^2\mathbf{I}_{\mathbf{M_i}}$. What is more, $\mathbf{H} = \sigma_v^2$. Hence,

$$Cov_\xi(Y_{idj}, Y_{i'd'j'}) = \begin{cases} 0 & \text{for} \quad i \neq i' \vee d \neq d' \\ \sigma_e^2 + x_{ij}^2\sigma_v^2 & \text{for} \quad i = i' \wedge j = j' \\ x_{ij}x_{i'j'}\sigma_v^2 & \text{for} \quad i = i' \wedge d = d' \wedge j \neq j' \end{cases} \tag{5}$$

The second model, which will be called special case 2, is nested error regression model similar to the one proposed by Battese, Harter and Fuller (1988) (but they do not consider longitudinal model, and they consider domain specific v random component):

$$Y_{idj} = \mathbf{x_{idj}}\boldsymbol{\beta}_d + v_{id} + e_{idj} \tag{6}$$

where $\mathbf{x_{idj}} = \begin{bmatrix} x_{idj1} & x_{idj2} & \dots & x_{idjp} \end{bmatrix}$

In the considered model we assume that $\mathbf{R_{id}} = \sigma_e^2\mathbf{I}_{\mathbf{M_i}}$. What is more, $\mathbf{H} = \sigma_v^2$. Hence, $Cov_\xi(Y_{idj}, Y_{i'd'j'})$ is given by (5) where $\forall_i \forall_j \; x_{ij} = 1$.

We have assumed that the population may change in time and that one population element may change its domain affiliation in time. Observations of new element of the population or observations of the population element after the change of the domain affiliation are treated as realizations of new profile (3). Hence, because of covariance structure where nonzero covariances are only within profiles, we assume independence of observations for some population element before and after changing domain affiliation.

II. PSEUDO-EBLUP UNDER SPECIAL CASE 1

Prasad and Rao (1999) propose to use EBLUP based on aggregated version (inclusion probabilities are included) of the unit level model and call the resulting predictor Pseudo-EBLUP. Another pseudo-EBLUP is presented by You and Rao (2002). These predictors in the case of survey conducted in one period are studied by Bleuer, Godbout and Morin (2007). In the above mentioned papers Psudo-EBLUP is used for data from single period.

Let us consider longitudinal survey. Let π_{ij} be inclusion probability of the i-th population element (which belongs to the d-th domain) in the period j. Hence, the Horvitz-Thompson estimator of the d-th domain mean in the period j is given by:

$$\hat{\theta}_{dj} = \sum\nolimits_{i \in s_{dj}} \frac{Y_{idj}}{\pi_{ij}} \left(\sum\nolimits_{i \in s_{dj}} \frac{1}{\pi_{ij}} \right)^{-1} = \sum\nolimits_{i \in s_{dj}} w_{ij} Y_{idj} , \tag{7}$$

where $w_{ij} = \frac{1}{\pi_{ij}} \left(\sum\nolimits_{i \in s_{dj}} \frac{1}{\pi_{ij}} \right)^{-1}$. Then, we transform type B model (unit level model) (4), into type A model (area level model):

$$\sum\nolimits_{i \in s_{dj}} w_{ij} Y_{idj} = \sum\nolimits_{i \in s_{dj}} w_{ij} \left(\beta_d x_{ij} + v_{id} x_{ij} + e_{idj} \right), \tag{8}$$

and hence

$$\hat{\theta}_{dj} = \beta_d \hat{\bar{x}}_{dj} + \ddot{v}_{dj} + \ddot{e}_{dj} , \tag{9}$$

where $\hat{\bar{x}}_{dj} = \sum\nolimits_{i \in s_{dj}} w_{ij} x_{ij}$, $\ddot{v}_{dj} = \sum\nolimits_{i \in s_{dj}} w_{ij} x_{ij} v_{id}$ and $\ddot{e}_{dj} = \sum\nolimits_{i \in s_{dj}} w_{ij} e_{idj}$, $\ddot{e}_{dj}$ are independent random components with variances $\sigma_e^2 \sum\nolimits_{i \in s_{dj}} w_{ij}^2$ and vectors $\ddot{\mathbf{v}}_{\mathbf{d}} = \left[\ddot{v}_{d1} \ \ldots \ \ddot{v}_{dj} \ \ldots \ \ddot{v}_{dm_d} \right]^T$ (where d=1,...,D and m_d is number of periods when at least one element of the *d*-th domain is observed) are independent random vectors with zero vector of expected values and variance-covariance matrix (under aggregated model (9)) $D_{\xi}^{2}(\ddot{\mathbf{v}}_{\mathbf{d}}) = \mathbf{G}_{\mathbf{d}} = \sigma_{v}^{2}\cancel{\mathbf{G}}_{\mathbf{d}}$, where

$$\cancel{\mathbf{G}}_{\mathbf{d}} = \begin{bmatrix} \sum_{i\in s_{d1}} w_{i1}^{2}x_{i1}^{2} & \dots & \sum_{i\in s_{d1}\cap s_{dj}} w_{i1}w_{ij}x_{i1}x_{ij} & \dots & \sum_{i\in s_{d1}\cap s_{dj}} w_{i1}w_{im_d}x_{i1}x_{im_d} \\ \dots & \dots & \dots & \dots & \dots \\ \sum_{i\in s_{d1}\cap s_{dj}} w_{ij}w_{i1}x_{ij}x_{i1} & \dots & \sum_{i\in s_{dj}} w_{ij}^{2}x_{ij}^{2} & \dots & \sum_{i\in s_{d1}\cap s_{dj}} w_{ij}w_{im_d}x_{ij}x_{im_d} \\ \dots & \dots & \dots & \dots & \dots \\ \sum_{i\in s_{d1}\cap s_{dj}} w_{im_d}w_{i1}x_{im_d}x_{i1} & \dots & \sum_{i\in s_{d1}\cap s_{dj}} w_{im_d}w_{ij}x_{im_d}x_{ij} & \dots & \sum_{i\in s_{dm_d}} w_{im_d}^{2}x_{im_d}^{2} \end{bmatrix} \tag{10}$$

Hence (based on Henderson's theorem, 1950), BLUP of domain total for aggregated model (4) called pseudo-BLUP is given by:

$$\hat{\theta}_{d^*j^*}^{Pseudo-BLUP} = N_{d^*j^*}\left(\hat{\hat{x}}_{d^*j^*}\hat{\beta}_{d} + \sigma_{v}^{2}\cancel{\mathbf{g}}_{d^*j^*}\mathbf{V}_{(aggr)\mathbf{dss}}^{-1}(\hat{\boldsymbol{\theta}}_{\mathbf{d}} - \hat{\hat{\mathbf{x}}}_{\mathbf{d}}\hat{\beta}_{d})\right), \tag{11}$$

where $\mathbf{V}_{(aggr)\mathbf{dss}} = \sigma_{e}^{2}diag_{1\le j\le m_d}\left(\sum_{i\in s_{dj}} w_{ij}^{2}\right) + \sigma_{v}^{2}\cancel{\mathbf{G}}_{\mathbf{d}}$, $\cancel{\mathbf{g}}_{d^*j^*}$ is j*-th row of $\cancel{\mathbf{G}}_{\mathbf{d}^*}$ matrix,

$$\hat{\boldsymbol{\theta}}_{\mathbf{d}} = \begin{bmatrix} \hat{\theta}_{d1} & \dots & \hat{\theta}_{dj} & \dots & \hat{\theta}_{dm_d} \end{bmatrix}^{T}, \hat{\hat{\mathbf{x}}}_{\mathbf{d}} = \begin{bmatrix} \hat{\hat{x}}_{d1} & \dots & \hat{\hat{x}}_{dj} & \dots & \hat{\hat{x}}_{dm_d} \end{bmatrix}^{T},$$

$$\hat{\beta}_{d} = (\hat{\hat{\mathbf{x}}}_{\mathbf{d}}^{T}\mathbf{V}_{(aggr)\mathbf{dss}}^{-1}\hat{\hat{\mathbf{x}}}_{\mathbf{d}})^{-1}\hat{\hat{\mathbf{x}}}_{\mathbf{d}}^{T}\mathbf{V}_{(aggr)\mathbf{dss}}^{-1}\hat{\boldsymbol{\theta}}_{\mathbf{d}},$$

Hence, the MSE of Pseudo-BLUP is given by

$$MSE_{\xi}(\hat{\theta}_{BLU}^{s}) = Var_{\xi}(\hat{\theta}_{BLU}^{s} - \theta') = g_{1}^{s}(\boldsymbol{\delta}) + g_{2}^{s}(\boldsymbol{\delta}) \tag{12}$$

where

$$g_{1}^{s}(\boldsymbol{\delta}) = N_{d^*j^*}^{2}\left(\sigma_{v}^{2}\sum_{i\in s_{d^*j^*}} w_{ij^*}^{2}x_{ij^*}^{2} - \sigma_{v}^{4}\cancel{\mathbf{g}}_{d^*j^*}\mathbf{V}_{(aggr)\mathbf{d^*ss}}^{-1}\cancel{\mathbf{g}}_{d^*j^*}^{T}\right), \tag{13}$$

$$g_{2}^{s}(\boldsymbol{\delta}) = N_{d^*j^*}^{2}\left(\hat{\hat{x}}_{d^*j^*} - \sigma_{v}^{2}\cancel{\mathbf{g}}_{d^*j^*}\mathbf{V}_{(aggr)\mathbf{dss}}^{-1}\hat{\hat{\mathbf{x}}}_{\mathbf{d}}\right)^{2}(\hat{\hat{\mathbf{x}}}_{\mathbf{d}}^{T}\mathbf{V}_{(aggr)\mathbf{dss}}^{-1}\hat{\hat{\mathbf{x}}}_{\mathbf{d}})^{-1}, \tag{14}$$

If the unknown parameters in (11) are replaced by their estimates predictor called Pseudo-EBLUP is obtained. Its MSE (using theorem of Datta and Lahiri (2000)) is given by:

$$MSE_{\xi}(\hat{\theta}^{S}_{EBLU}(\hat{\boldsymbol{\delta}})) = g_1^s(\boldsymbol{\delta}) + g_2^s(\boldsymbol{\delta}) + g_3^{s*}(\boldsymbol{\delta}) + o(D^{-1}) \tag{15}$$

where $g_1^s(\boldsymbol{\delta})$ and $g_2^s(\boldsymbol{\delta})$ are given by (13) and (14) respectively and

$$g_3^{s*}(\boldsymbol{\delta}) = N^2_{d*j*}\left(q_{ee}I^{(-1)}_{vv} + 2q_{ev}I^{(-1)}_{ve} + q_{vv}I^{(-1)}_{ee}\right) \tag{16}$$

where

$$q_{ee} = \sigma_v^4 \mathbf{g}_{d*j*}\mathbf{V}^{-1}_{(aggr)\mathbf{dss}} diag_{1\le j\le m_d}\left(\sum\nolimits_{i\in s_{dj}} w_{ij}^2\right)\mathbf{V}^{-1}_{(aggr)\mathbf{dss}} diag_{1\le j\le m_d}\left(\sum\nolimits_{i\in s_{dj}} w_{ij}^2\right)\mathbf{V}^{-1}_{(aggr)\mathbf{dss}}\mathbf{g}^T_{d*j*} \tag{17}$$

$$q_{vv} = \mathbf{g}_{d*j*}(\mathbf{I}_{\mathbf{m_d}} - \sigma_v^2\mathbf{V}^{-1}_{(aggr)\mathbf{dss}}\mathbf{G}_{\mathbf{d}})\mathbf{V}^{-1}_{(aggr)\mathbf{dss}}(\mathbf{I}_{\mathbf{m_d}} - \sigma_v^2\mathbf{V}^{-1}_{(aggr)\mathbf{dss}}\mathbf{G}_{\mathbf{d}})^T\,\mathbf{g}^T_{d*j*} \tag{18}$$

$$q_{ev} = -\sigma_v^2 \mathbf{g}_{d*j*}\mathbf{V}^{-1}_{(aggr)\mathbf{dss}} diag_{1\le j\le m_d}\left(\sum\nolimits_{i\in s_{dj}} w_{ij}^2\right)\mathbf{V}^{-1}_{(aggr)\mathbf{dss}}(\mathbf{I}_{\mathbf{m_d}} - \sigma_v^2\mathbf{V}^{-1}_{(aggr)\mathbf{dss}}\mathbf{G}_{\mathbf{d}})^T\,\mathbf{g}^T_{d*j*} \tag{19}$$

$$\mathbf{I}_{\delta}^{-1} = \begin{bmatrix} I^{(-1)}_{vv} & I^{(-1)}_{ve} \\ I^{(-1)}_{ve} & I^{(-1)}_{ee} \end{bmatrix} \tag{20}$$

$$I^{(-1)}_{vv} = 2b^{-1}\sum_{d=1}^{D}\sum_{i=1}^{n_d} b_{id}^{-2}\left(\sum_{j=1}^{m_{id}} x_{ij}^2\right)^2,\ I^{(-1)}_{ve} = -2b^{-1}\sum_{d=1}^{D}\sum_{i=1}^{n_d} b_{id}^{-2}\left(\sum_{j=1}^{m_{id}} x_{ij}^2\right),$$

$$I^{(-1)}_{ee} = 2b^{-1}\sum_{d=1}^{D}\sum_{i=1}^{n_d}\left((m_{id}-1)\sigma_e^{-4} + b_{id}^{-2}\right),\ b_{id} = \sigma_e^2 + \sigma_v^2\sum_{j=1}^{m_{id}} x_{ij}^2,$$

$$b = \left(\sum_{d=1}^{D}\sum_{i=1}^{n_d}\left((m_{id}-1)\sigma_e^{-4} + b_{id}^{-2}\right)\right)\left(\sum_{d=1}^{D}\sum_{i=1}^{n_d} b_{id}^{-2}\left(\sum_{j=1}^{m_{id}} x_{ij}^2\right)^2\right) - \left(\sum_{d=1}^{D}\sum_{i=1}^{n_d} b_{id}^{-2}\left(\sum_{j=1}^{m_{id}} x_{ij}^2\right)\right)^2$$

The MSE estimator of the pseudo-EBLUP when unknown parameters are replaced by its Restricted Maximum Likelihood (REML) estimates is given by

$$M\hat{S}E_{\xi}\left(\hat{\theta}^{S}_{EBLU}(\hat{\boldsymbol{\delta}})\right) = g_1^s(\hat{\boldsymbol{\delta}}) + g_2^s(\hat{\boldsymbol{\delta}}) + 2g_3^s(\hat{\boldsymbol{\delta}}) - \mathbf{B}^T_{\hat{\boldsymbol{\delta}}}(\hat{\boldsymbol{\delta}})\frac{\partial g_1^s(\hat{\boldsymbol{\delta}})}{\delta\boldsymbol{\delta}}, \tag{21}$$

where vector of biases $\mathbf{B}^T_{\hat{\boldsymbol{\delta}}}(\boldsymbol{\delta})$ for REML estimators assuming normality is given by:

$$\mathbf{B}_{\hat{\boldsymbol{\delta}}^{REML}}(\boldsymbol{\delta}) = o(D^{-1}), \tag{22}$$

and hence the last element in (21) is omitted, $g_1^s(\hat{\boldsymbol{\delta}}), g_2^s(\hat{\boldsymbol{\delta}}), g_3^s(\hat{\boldsymbol{\delta}})$ are given by (13), (14), (16) respectively, where vector of unknown parameters $\boldsymbol{\delta}$ is replaced by estimator $\hat{\boldsymbol{\delta}}$.

The MSE estimator of the pseudo-EBLUP when unknown parameters are replaced by their Maximum Likelihood (ML) estimates is given by (21), where $g_1^s(\hat{\boldsymbol{\delta}}), g_2^s(\hat{\boldsymbol{\delta}}), g_3^s(\hat{\boldsymbol{\delta}})$ are given by (13), (14), (16) respectively, where vector of unknown parameters $\boldsymbol{\delta}$ is replaced by estimator $\hat{\boldsymbol{\delta}}$, elements of $\frac{\partial g_1^s(\boldsymbol{\delta})}{\delta\boldsymbol{\delta}}$ are given by:

$$\frac{\partial g_1^s(\boldsymbol{\delta})}{\partial\sigma_e^2} = -N^2_{d^*j^*}\sigma_v^4\mathbf{g}_{d^*j^*}\mathbf{V}^{\mathbf{-1}}_{(aggr)\mathbf{d^*ss}}diag_{1\le j\le m_d}\left(\sum\nolimits_{i\in s_{dj}} w_{ij}^2x_{ij^*}^2\right)\mathbf{V}^{\mathbf{-1}}_{(aggr)\mathbf{d^*ss}}\mathbf{g}^T_{d^*j^*}$$

$$\frac{\partial g_1^s(\boldsymbol{\delta})}{\partial\sigma_v^2} = N^2_{d^*j^*}\left(\sum\nolimits_{i\in s_{d^*j^*}} w_{ij}^2x_{ij^*}^2 - 2\sigma_v^2\mathbf{g}_{d^*j^*}\mathbf{V}^{\mathbf{-1}}_{(aggr)\mathbf{d^*ss}}\mathbf{g}^T_{d^*j^*} + \sigma_v^4\mathbf{g}_{d^*j^*}\mathbf{V}^{\mathbf{-1}}_{(aggr)\mathbf{d^*ss}}\mathbf{G}_{\mathbf{d^*}}\mathbf{V}^{\mathbf{-1}}_{(aggr)\mathbf{d^*ss}}\mathbf{g}^T_{d^*j^*}\right)$$

where the vector of bias of ML estimators is given by

$$\mathbf{B}_{\hat{\boldsymbol{\delta}}^{ML}}(\boldsymbol{\delta}) = \frac{1}{2}\mathbf{I}_{\delta}^{-1}(\delta)col_{1\le k\le D}tr\left[\mathbf{I}_{\beta}^{-1}(\delta)\frac{\partial}{\partial\delta_k}\mathbf{I}_{\beta}(\delta)\right] + o(D^{-1}) \tag{23}$$

where $\mathbf{I}_{\delta}^{-1}$ is given by (20) and

$$col_{1\le k\le q}tr\left[\mathbf{I}_{\beta}^{-1}(\boldsymbol{\delta})\frac{\partial}{\partial\delta_k}\mathbf{I}_{\beta}(\boldsymbol{\delta})\right] =$$

$$= -\left[\sum_{d=1}^{D}\left(\sum_{i=1}^{n_d}b_{id}^{-1}\sum_{j=1}^{m_{id}}x_{ij}^2\right)^{-1}\left(\sum_{i=1}^{n_d}b_{id}^{-2}\sum_{j=1}^{m_{id}}x_{ij}^2\right)\quad\sum_{d=1}^{D}\left(\sum_{i=1}^{n_d}b_{id}^{-1}\sum_{j=1}^{m_{id}}x_{ij}^2\right)^{-1}\left(\sum_{i=1}^{n_d}b_{id}^{-2}\left(\sum_{j=1}^{m_{id}}x_{ij}^2\right)^2\right)\right]^T$$

III. PSEUDO-EBLUP UNDER SPECIAL CASE 2

Based on Henderson's theorem (1950), the BLUP of the domain total for aggregated version of model (6) called the pseudo-BLUP is given by:

$$\hat{\theta}_{d^*j^*}^{Pseudo-BLUP} = N_{d^*j^*}\left(\hat{\hat{\mathbf{x}}}_{\mathbf{d^*j^*}}\hat{\boldsymbol{\beta}}_{\mathbf{d}} + \sigma_v^2 \mathbf{g}_{d^*j^*}\mathbf{V}_{(aggr)\mathbf{dss}}^{-1}(\hat{\boldsymbol{\theta}}_{\mathbf{d}} - \hat{\hat{\mathbf{X}}}_{\mathbf{d}}\hat{\boldsymbol{\beta}}_{\mathbf{d}})\right) \tag{24}$$

where $\mathbf{V}_{(aggr)\mathbf{dss}} = \sigma_e^2 diag_{1\le j\le m_d}\left(\sum_{i\in s_{dj}} w_{ij}^2\right) + \sigma_v^2 \mathbf{G}_{\mathbf{d}}$, $\mathbf{g}_{d^*j^*}$ is j*-th row of $\mathbf{G}_{\mathbf{d^*}}$ matrix given in this case by (10), where $\forall_{i,j} x_{ij} = 1$, $\hat{\boldsymbol{\theta}}_{\mathbf{d}} = \begin{bmatrix} \hat{\theta}_{d1} & \ldots & \hat{\theta}_{dj} & \ldots & \hat{\theta}_{dm_d} \end{bmatrix}^T$, $\hat{\hat{\mathbf{X}}}_{\mathbf{d}} = \begin{bmatrix} \hat{\hat{\mathbf{x}}}_{\mathbf{d1}}^T & \ldots & \hat{\hat{\mathbf{x}}}_{\mathbf{dj}}^T & \ldots & \hat{\hat{\mathbf{x}}}_{\mathbf{dm_d}}^T \end{bmatrix}^T$ and $\hat{\boldsymbol{\beta}}_{\mathbf{d}} = (\hat{\hat{\mathbf{X}}}_{\mathbf{d}}^T \mathbf{V}_{\mathbf{dss}}^{-1} \hat{\hat{\mathbf{X}}}_{\mathbf{d}})^{-1} \hat{\hat{\mathbf{X}}}_{\mathbf{d}}^T \mathbf{V}_{\mathbf{dss}}^{-1} \hat{\boldsymbol{\theta}}_{\mathbf{d}}$

For aggregated version of model (6), equations (12), (15), (21), (23) are still valid, but:

$$g_1^s(\boldsymbol{\delta}) = N_{d^*j^*}^2\left(\sigma_v^2 \sum_{i\in s_{d^*j^*}} w_{ij^*}^2 - \sigma_v^4 \mathbf{g}_{d^*j^*}\mathbf{V}_{(aggr)\mathbf{d^*ss}}^{-1}\mathbf{g}_{d^*j^*}^T\right), \tag{25}$$

$$g_2^s(\boldsymbol{\delta}) = N_{d^*j^*}^2\left(\hat{\hat{\mathbf{x}}}_{\mathbf{dj}} - \sigma_v^2 \mathbf{g}_{d^*j^*}\mathbf{V}_{(aggr)\mathbf{dss}}^{-1}\hat{\hat{\mathbf{X}}}_{\mathbf{d}}\right)^2 (\hat{\hat{\mathbf{X}}}_{\mathbf{d}}^T \mathbf{V}_{\mathbf{dss}}^{-1} \hat{\hat{\mathbf{X}}}_{\mathbf{d}})^{-1}, \tag{26}$$

$$g_3^{s*}(\boldsymbol{\delta}) = N_{d^*j^*}^2\left(q_{ee}I_{vv}^{(-1)} + 2q_{ev}I_{ve}^{(-1)} + q_{vv}I_{ee}^{(-1)}\right) \tag{27}$$

where q_{ee}, q_{ev}, q_{vv} are given by (17), (18), (19) but where $\mathbf{G}_{\mathbf{d}}$ is given by (10), where $\forall_{i,j} x_{ij} = 1$, and $I_{vv}^{(-1)} = 2b^{-1}\sum_{d=1}^{D}\sum_{i=1}^{n_d} b_{id}^{-2} m_{id}^2$, $I_{ve}^{(-1)} = -2b^{-1}\sum_{d=1}^{D}\sum_{i=1}^{n_d} b_{id}^{-2} m_{id}$, $I_{ee}^{(-1)} = 2b^{-1}\sum_{d=1}^{D}\sum_{i=1}^{n_d}\left((m_{id}-1)\sigma_e^{-4} + b_{id}^{-2}\right)$, $b_{id} = \sigma_e^2 + \sigma_v^2 m_{id}$,

$$b = \left(\sum_{d=1}^{D}\sum_{i=1}^{n_d}\left((m_{id}-1)\sigma_e^{-4} + b_{id}^{-2}\right)\right)\left(\sum_{d=1}^{D}\sum_{i=1}^{n_d} b_{id}^{-2} m_{id}^2\right) - \left(\sum_{d=1}^{D}\sum_{i=1}^{n_d} b_{id}^{-2} m_{id}\right)^2$$

What is more,

$$\frac{\partial g_1^s(\boldsymbol{\delta})}{\partial \sigma_e^2} = -N_{d^*j^*}^2 \sigma_v^4 \mathbf{g}_{d^*j^*}\mathbf{V}_{(aggr)\mathbf{d^*ss}}^{\mathbf{-1}} diag_{1\le j\le m_d}\left(\sum_{i\in s_{dj}} w_{ij}^2\right)\mathbf{V}_{(aggr)\mathbf{d^*ss}}^{\mathbf{-1}}\mathbf{g}_{d^*j^*}^T$$

$$\frac{\partial g_1^s(\boldsymbol{\delta})}{\partial \sigma_v^2} = N_{d*j*}^2 \left(\sum\nolimits_{i \in s_{d*j*}} w_{ij}^2 - 2\sigma_v^2 \mathbf{g}_{d*j*} \mathbf{V}^{\mathbf{-1}}_{(aggr)\mathbf{d*ss}} \mathbf{g}_{d*j*}^T + \sigma_v^4 \mathbf{g}_{d*j*} \mathbf{V}^{\mathbf{-1}}_{(aggr)\mathbf{d*ss}} \mathbf{G}_{\mathbf{d*}} \mathbf{V}^{\mathbf{-1}}_{(aggr)\mathbf{d*ss}} \mathbf{g}_{d*j*}^T \right)$$

$$col_{1 \le k \le q} tr \left[\mathbf{I}_\beta^{-1}(\boldsymbol{\delta}) \frac{\partial}{\partial \delta_k} \mathbf{I}_\beta(\boldsymbol{\delta}) \right] =$$

$$= - \left[\sum_{d=1}^{D} tr \left(\left(\sum_{i=1}^{n_d} b_{id}^{-1} \mathbf{X}_{\mathbf{sid}}^{\mathbf{T}} \mathbf{X}_{\mathbf{sid}} \right)^{-1} \left(\sum_{i=1}^{n_d} b_{id}^{-2} \mathbf{X}_{\mathbf{sid}}^{\mathbf{T}} \mathbf{X}_{\mathbf{sid}} \right) \right) \; \sum_{d=1}^{D} tr \left(\left(\sum_{i=1}^{n_d} b_{id}^{-1} \mathbf{X}_{\mathbf{sid}}^{\mathbf{T}} \mathbf{X}_{\mathbf{sid}} \right)^{-1} \left(\sum_{i=1}^{n_d} m_{id} b_{id}^{-2} \mathbf{X}_{\mathbf{sid}}^{\mathbf{T}} \mathbf{X}_{\mathbf{sid}} \right) \right) \right]^T$$

where $\mathbf{X}_{\mathbf{sid}}$ is $m_{id} \times p$ matrix of auxiliary variables.

IV. SIMULATION ANALYSES

The Monte Carlo simulation analyses are based on real data on *N*=314 Polish poviats (excluding cites with poviat's rights), what is NTS 4 level, for *M*=4 years 2005-2008 (from www.stat.gov.pl). The problem is to estimate subpopulations (domains) totals for *D*=6 regions (NTS 1 level) in 2008. The variable of interest is poviats' own incomes (in PLN) and the auxiliary variable is the population size in poviats (in persons). Two simulations are conducted using R (R Development Core Team, 2010). In the simulations accuracy of the Pseudo-EBLUP will be compared with accuracy of two calibration estimators (Rao (2003) pp. 17-18) which will be denoted by GREG1 and GREG2 . Both calibration estimators are of the form $\hat{t}_{d*j*}^{GREG} = \sum_{i \in s_{d*j*}} w_{sij*} y_{ij*}$, but weights w_{sij*} are solutions for GREG1 of $\begin{cases} \sum_{i \in s_{d*j*}} \dfrac{\left(w_{sij*} - d_i\right)^2}{d_i q_i} \to \min \\ \sum_{i \in s_{d*j*}} w_{sij*} \mathbf{x_i} = \sum_{i \in \Omega_{d*j*}} \mathbf{x_i} \end{cases}$ and for GREG2 of

$$\begin{cases} \sum_{i \in s_{j*}} \dfrac{\left(w_{sij*}^{(2)} - d_i\right)^2}{d_i q_i} \to \min \\ \sum_{i \in s_{j*}} w_{sij*}^{(2)} \mathbf{x_i} = \sum_{i \in \Omega_{j*}} \mathbf{x_i} \end{cases} .$$

The first simulation is design-based. In this case sample of size *n*=79 elements (c.a. 25% of population size) is balanced panel sample, which is drawn at random in the first period with inclusion probabilities proportional to the value of the auxiliary variable in this period. For this sample size it was possible to estimate all of domain totals in each iteration even using direct estimators. The number of samples drawn in the simulation equals 10 000.

The second simulation is model based. In this case one sample is drawn using sample design described above what gives the division if the population on the sampled and unsampled part. Than 10 000 populations are generated using model (6) (with one auxiliary variable and constant) with parameters computed based on real population data and random components with the following distributions: in the model denoted in the simulation as NN case – normal distribution of both v_{id} and e_{idj}, UU – uniform distribution of both v_{id} and e_{idj}, EE – shifted exponential distribution of both v_{id} and e_{idj}. What is more, to study the problem of model misspecification, 10 000 population are generated based on modified model (6), where instead of the auxiliary variable its natural logarithm is used, where random components have the following distributions: NNm case – normal distributions, UUm – uniform distributions, EEm – shifted exponential distributions.

Selected results are presented in the table 1. Design biases of all of estimators/predictors are around zero. The design accuracy of the Pseudo-EBLUP (P-EBLUP) is on average better than the design accuracy of GREG2 but on average worse than the design accuracy of GREG1. The results of model-based simulations are similiar for all of models (hence, in the table 1 results for one model are presented – EEm model). On average over domains, the smallest values of absolute model biases and the smallest values of model-RMSEs for all of models are observed for P-EBLUP.

Table 1. Selected results of the Monte Carlo simulations

Result:	Estimator:	Domain:					
		1	2	3	4	5	6
Relative design-biases (in %)	GREG1	–0,57	0,58	–0,31	–0,78	0,92	–0,26
	GREG2	0,30	–0,70	0,06	0,03	–0,25	0,52
	P-EBLUP	2,94	2,48	–0,09	–2,92	0,91	–0,96
Relative design-RMSE (in %)	GREG1	8,67	5,65	4,45	5,71	9,24	5,28
	GREG2	21,43	20,33	18,98	21,10	27,22	23,27
	P-EBLUP	18,78	18,91	18,40	12,50	23,11	15,55
Relative model-biases (in %) for EEm model	GREG1	33,05	22,90	32,08	34,26	13,59	23,98
	GREG2	41,67	9,82	43,10	25,98	–2,76	39,05
	P-EBLUP	4,47	1,17	8,66	–13,52	–2,51	1,97
Relative model-RMSE (in %) for EEm model	GREG1	33,40	23,42	32,82	34,74	15,18	25,08
	GREG2	41,95	10,99	43,66	26,61	7,29	39,74
	P-EBLUP	10,77	8,10	16,67	18,15	13,91	14,30

V. CONCLUSION

In the paper Pseudo-EBLUP is proposed for longitudinal data. Its design- and model-accuracy is studied in Monte Carlo simulation analysis based on real data. The prediction accuracy of the predictor is on average better in the simulation comparing with both considered calibration estimators but values of design MSEs are on average between the design MSEs of the considered calibration estimators.

REFERENCES

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

Bleuer S.R, Godbout S., Morin Y. (2007), *Evaluation of small domain estimators for the survey of employment payroll and hours*, SSC Annual Meeting, Proceedings of the Survey Methods Section

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.

Depmster A.P., Rubin D.B. and Tsutakawa R.K. (1981), Estimation in covariance components models, *Journal of the American Statistical Association,* Vol. 76, No. 374, pp. 341– 353

Deville, J.C., Särndal, C.E. (1992), Calibration estimators in survey sampling, *Journal of the American Statistical Association*, 87, 376–382.

Hedeker, D., Gibbons, R.D. (2006), *Longitudinal Data Analysis*, John Wiley & Sons, New Jersey.

Henderson C.R. (1950). Estimation of genetic parameters (Abstract). *Annals of Mathematical Statistics*, 21, 309–310.

Moura, F.A.S. and Holt, D. (1999). Small area estimation using multilevel models. *Survey Methodology*, 25, 73–80.

Prasad N.G.N and Rao J.N.K. (1990). The estimation of mean the mean squared error of small area estimators. *Journal of the American Statistical Association*, 85, 163–171.

Prasad N.G.N and Rao J.N.K. (1999). On robust small area estimation using a simple random effects model, *Survey Methodology, 25,* 67–72

R Development Core Team, 2010. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna.

Rao, J.N.K. (2003), *Small area estimation*. John Wiley & Sons, New York.

Verbeke G., Molenberghs G. (2000), *Linear Mixed Models for Longitudinal Data*, Springer-Verlag, New York.

You Y., Rao J.N.K. (2002), A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights, *The Canadian Journal of Statistics*, 30, 431–439

Tomasz Żądło

O PREDYKTORZE PSEUDO-EBLUP DLA PEWNEGO MODELU NADPOPULACJI ZE ZMIENNYMI DODATKOWYMI DLA DANYCH WIELOOKRESOWYCH

Rozważany jest problem modelowania profili wielookresowych zakładając, że zarówno populacja jak i przynależność elementów populacji do podpopulacji może zmieniać się w czasie. Dla danych przekrojowo-czasowych zakładamy pewien model mieszany ze składnikami losowymi

specyficznymi dla podpopulacji i elementów populacji (por. Verbeke, Molenberghs, 2000; Hedeker, Gibbons, 2006), który jest przypadkiem szczególnym ogólnego mieszanego modelu liniowego. Zostaną przedstawione pseudo-empiryczne najlepsze liniowe nieobciążone predyktory wynikające z podejścia mieszanego (wspomaganie modelami nadpopulacji), ich błędy średniokwadratowe i ich estymatory. W badaniu symulacyjnym ich dokładność zostanie porównana z pewnymi estymatorami kalibrowanymi również wynikającymi z podejścia mieszanego.