ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 285, 2013

Tomasz Żądło[*]

ON PARAMETER ESTIMATION OF SOME LONGITUDINAL MODEL

Abstract. The problem of modeling longitudinal profiles is considered assuming that the population and elements' affiliation to subpopulations may change in time. Some longitudinal model which is a special case of the general linear model (GLM) and the general linear mixed model (GLMM) is studied. In the model two random components are included under assumptions of simultaneous spatial autoregressive process (SAR) and temporal first-order autoregressive process (AR(1)) respectively. The accuracy of model parameters' restricted maximum likelihood estimators is considered in the simulation.

Key words: longitudinal data, restricted maximum likelihood, MSE.

I. INTRODUCTION

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 subpopulations (domains) 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 subpopulation elements for which observations are available in the period t is denoted by s_{dt} and its size by n_{dt}. Let: $\Omega_{rdt} = \Omega_{dt} - s_{dt}$, $N_{rdt} = N_{dt} - n_{dt}$.

Let M_{id} denotes the number of periods when the i-th population element 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}. Let $m_{rid} = M_{id} - m_{id}$. It is assumed 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).

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

It means that i and t completely identify domain affiliation but additional subscript d will be needed as well. The set of elements which belong at least in one of periods t=1,...,M to sets Ω_t is denoted by Ω and its size by N. Similarly, sets Ω_d, s, s_d, Ω_{rd} of sizes N_d, n, n_d, N_{rd} respectively are defined as sets of elements which belong at least in one of periods t=1,...,M to sets Ω_{dt}, s_t, s_{dt}, Ω_{rdt} respectively. The d*-th domain of interest in the period of interest t^* will be denoted by Ω_{d*t*}, and the set of elements which belong at least in one of periods t=1,...,M to sets Ω_{d*t*} will be denoted by Ω_{d*}. The introduced notations allow to assume that the domain affiliations of population elements change in time.

II.LONGITUDINAL MODEL

In the small area estimation literature the problem of spatial correlation is studied but for one period including both area-level models (Molina, Salvati and Pratesi, 2009; Petruci and Salvati, 2006; Petruci, Pratesi and Salvati, 2005; Pratesi and Salvati, 2008) and unit-level models (Chandra, Salvati and Chambers, 2007, Salvati, Pratesi, Tzavidis and Chambers, 2009). In this paper superpopulation models used for longitudinal data (compare Verbeke and Molenberghs, 2000; Hedeker and Gibbons, 2006) are considered both with spatial and temporal correlation which are – what is important for further considerations – special cases of the General Linear Model (GLM) and the General Linear Mixed Model (GLMM). We propose the following model:

$$\mathbf{Y_d} = \mathbf{X_d}\boldsymbol{\beta}_\mathbf{d} + \mathbf{Z_d}\mathbf{v_d} + \mathbf{e_d}, \tag{1}$$

where $\mathbf{Y_d} = col_{1\le i\le N_d}(\mathbf{Y_{id}})$, where $\mathbf{Y_{id}}$ is a random vector, called profile, of size $M_{id}\times 1$, and $\mathbf{Y_d}$ (d=1,...,D) are assumed to be independent, $\mathbf{X_d} = col_{1\le i\le N_d}(\mathbf{X_{id}})$, where $\mathbf{X_{id}}$ is known matrix of size $M_{id}\times p$, $\mathbf{Z_d} = diag_{1\le i\le N_d}(\mathbf{Z_{id}})$, where $\mathbf{Z_{id}}$ is known vector of size $M_{id}\times 1$, $\mathbf{v_d} = col_{1\le i\le N_d}(v_{id})$, where v_{id} is a profile-specific random component and $\mathbf{v_d}$ (d=1,2...,D) are assumed to be independent, $\mathbf{e_d} = col_{1\le i\le N_d}(\mathbf{e_{id}})$, where $\mathbf{e_{id}}$ is a random component vector of size $M_{id} \times 1$ and $\mathbf{e_{id}}$ (*i*=1,...,*N*; *d*=1,...,*D*) are assumed to be independent, $\mathbf{v_d}$ and $\mathbf{e_d}$ are assumed to be independent.

What is more, it is assumed that vector of random components $\mathbf{v_d}$ obey assumptions of simultaneously spatial autoregressive (SAR) process:

$$\mathbf{v_d} = \rho_{(sp)}\mathbf{W}_d\mathbf{v_d} + \mathbf{u_d}, \tag{2}$$

where $\mathbf{W}_d$ is the spatial weight matrix for profiles $\mathbf{Y_{id}}$, $\mathbf{u_d} \sim (\mathbf{0}, \sigma_u^2\mathbf{I}_{N_d})$.
Hence,

$$\mathbf{v_d} \sim N\left(\mathbf{0}, \mathbf{R_d}\right), \tag{3}$$

where $\mathbf{R_d} = \sigma_u^2\mathbf{C_d^{-1}}$ and $\mathbf{C_d} = \left(\mathbf{I_{N_d}} - \rho_{(sp)}\mathbf{W_d}\right)\left(\mathbf{I}_{N_d} - \rho_{(sp)}\mathbf{W_d^T}\right)$.

Moreover, elements of $\mathbf{e_{id}}$ obey assumptions of autoregressive process AR(1):

$$e_{idj} = \rho_{(t)}e_{idj-1} + \varepsilon_{idj}. \tag{4}$$

Hence,

$$e_{\mathbf{id}} \sim N\left(\mathbf{0}, \mathbf{\Sigma_{id}}\right), \tag{5}$$

where elements of $\mathbf{\Sigma_{id}}$ are given by $\sigma_\varepsilon^2\rho_{(t)}^{|k-l|}\left(1-\rho_{(t)}^2\right)^{-1}$.

III. ESTIMATION OF PARAMETERS

The restricted maximum likelihood method (REML) was proposed by Thomson (1962) as written by Jiang (1996). What is important, the Gaussian REML is robust for nonnormality cases - as prooved by Jiang (1996) Gaussian REML estimators remain consistent and asymptotically normal even if normality does not hold.

Let $\mathbf{Y_{sd}} = col_{1\le i\le n_d}(\mathbf{Y_{sid}})$, where $\mathbf{Y_{sid}}$ is a random vector, called sample profile, of size $m_{id} \times 1$. Let

$$\mathbf{V}_{\mathbf{ss\,d}} = D_{\xi}^{2}(\mathbf{Y}_{\mathbf{ss\,d}}) = \sigma_u^2 \mathbf{Z}_{\mathbf{sd}} \mathbf{C}_{\mathbf{d}}^{\mathbf{-1}} \mathbf{Z}_{\mathbf{sd}}^{T} + diag_{1 \le i \le n_d}(\mathbf{\Sigma}_{\mathbf{ss\,id}}). \tag{6}$$

where $\mathbf{Z}_{\mathbf{sd}}$ is obtained from $\mathbf{Z}_{\mathbf{d}}$ by deleting rows for unsampled profiles, $\mathbf{\Sigma}_{\mathbf{ss\,id}}$ is a submatrix obtained from $\mathbf{\Sigma}_{\mathbf{id}}$ by deleting rows and columns for unsampled observations.

The restricted likelihood function for the considered model (1) is given by:

$$L = \prod_{d=1}^{D} \frac{1}{\sqrt{(2\pi)^n \det \mathbf{A}_{\mathbf{d}}^{\mathbf{T}} \mathbf{V}_{\mathbf{ss\,d}} \mathbf{A}_{\mathbf{d}}}} \exp\left\{ -\frac{1}{2} \mathbf{Y}_{\mathbf{sd}}^{\mathbf{T}} \left(\mathbf{A}_{\mathbf{d}}^{\mathbf{T}} \mathbf{V}_{\mathbf{ss\,d}} \mathbf{A}_{\mathbf{d}} \right)^{-1} \mathbf{Y}_{\mathbf{sd}} \right\} \tag{7}$$

where matrices $\mathbf{A}_{\mathbf{d}}$ (d=1,...,D) are any matrices of sizes $\left(\sum_{i=1}^{n_d} m_{id} \times 1 \right) \times \left(\left(\sum_{i=1}^{n_d} m_{id} \times 1 \right) - p \right)$ of rank $\left(\sum_{i=1}^{n_d} m_{id} \times 1 \right) - p$ such that: $\mathbf{A}_{\mathbf{d}}^{\mathbf{T}} \mathbf{X}_{\mathbf{sd}} = 0$. Matrices $\mathbf{A}_{\mathbf{d}}$ may be given by any $\left(\sum_{i=1}^{n_d} m_{id} \times 1 \right) - p$ linear independent rows of $(\mathbf{I} - \mathbf{X}_{\mathbf{sd}} (\mathbf{X}_{\mathbf{sd}}^{T} \mathbf{X}_{\mathbf{sd}})^{-1} \mathbf{X}_{\mathbf{sd}}^{T})$.

IV. SIMULATION ANALYSIS

Limited model-based simulation study prepared using R (R Development Core Team (2011)) is based on artificial data. Population of size N=200 elements is divided into D=10 domains of sizes {15, 15, 15, 20, 20, 20, 20, 25, 25, 25}. Number of periods M=3 and balanced panel sample is studied – in each period the same $n_d = 5$ elements from each domain are observed in the sample (overall sample size in each period is n=50). The purpose of the study is to predict D=10 domain totals for the last period.

Data are generated based on the model (1) where $\forall_{idj} x_{idj} = 1$, $\forall_{idj} z_{idj} = 1$, $\forall_d \beta_d = \beta$ and for arbitrary chosen values of parameters $\beta = 100$, $\sigma_\varepsilon^2 = 1$, $\sigma_u^2 = 1$. In the simulation the following values of $\rho_{(sp)}$ and $\rho_{(t)}$ are considered: 0,8; 0,3; -0,3 and -0,8 what gives sixteen pairs of these correlation coefficients (these pairs are presented on x-axis). Realizations of random components are generated using multivariate normal distribution. To maximize logarithm of the function (7) the *constrOpitm* R function was used.

For the assumed model (1) and under the assumptions made in the simulation (balanced panel sample, $\forall_{idj} x_{idj} = 1$, $\forall_{idj} z_{idj} = 1$, $\forall_d \beta_d = \beta$) the Best Linear Unbiased Predictor of the d**th* domain total in the t**th* period is given by

$$\hat{\theta}^{BLU}_{d*t*} = \sum_{i \in s_{d*t*}} Y_{id*t*} + N_{rd*t*}\hat{\mu} +$$

$$\boldsymbol{\gamma}^{\mathbf{T}}_{\mathbf{rd}*}\left(\sigma_u^2 \mathbf{Z}_{\mathbf{rd}*}\mathbf{C}^{\mathbf{-1}}_{\mathbf{d}*}\mathbf{Z}^{\mathbf{T}}_{\mathbf{sd}*}\right)\mathbf{V}^{-1}_{\mathbf{ss\,d}*}\left(\mathbf{Y}_{\mathbf{sd}*} - \mathbf{1}_{\sum_{i=1}^{n_{d*}} m_{id*}}\hat{\mu}\right), \qquad (8)$$

where $\hat{\mu} = \mathbf{1}^{\mathbf{T}}_{\sum_{d=1}^{D}\sum_{i=1}^{n_d} m_{id}} \mathbf{V}^{\mathbf{-1}}_{\mathbf{ss}}\mathbf{Y}_{\mathbf{s}} / \mathbf{1}^{\mathbf{T}}_{\sum_{d=1}^{D}\sum_{i=1}^{n_d} m_{id}} \mathbf{V}^{\mathbf{-1}}_{\mathbf{ss}}\mathbf{1}_{\sum_{d=1}^{D}\sum_{i=1}^{n_d} m_{id}}$, $\boldsymbol{\gamma}_{\mathbf{rd}*}$ is a $\sum_{i=1}^{n_{d*}} M_{rid*} \times 1$ vector of one's for observations in Ω_{rd*t*} and zero otherwise, $\mathbf{Z}_{\mathbf{sd}}$ and $\mathbf{Z}_{\mathbf{rd}}$ are obtained from $\mathbf{Z}_{\mathbf{d}} = diag_{1 \le i \le N_d}(\mathbf{1}_{M_{id}})$ by deleting rows for unsampled and sampled profiles respectively, $\mathbf{1}_a$ is $a \times 1$ vector of one's.

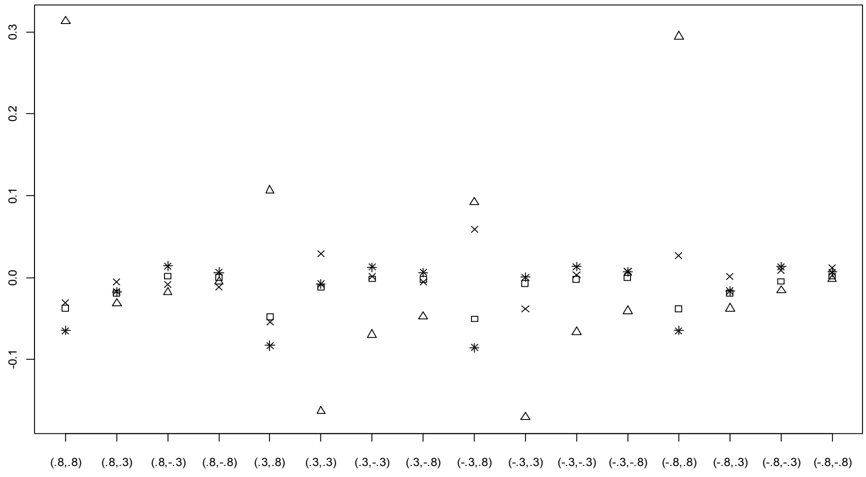

Figure 1. Absolute biases of estimators of: σ_e^2 - □, σ_u^2 - Δ, $\rho_{(sp)}$ - x, $\rho_{(t)}$ - *

Source: own work

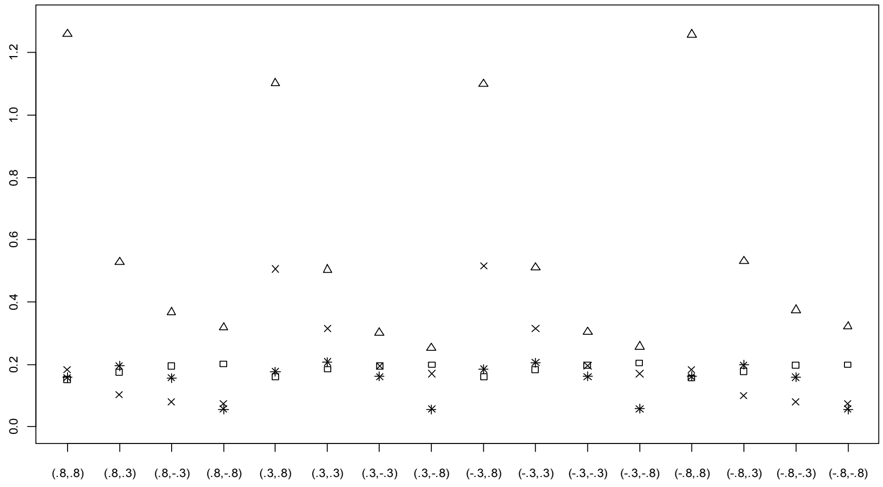

Figure 2. Absolute RMSEs of estimators of: σ_e^2 - □, σ_u^2 - Δ, $\rho_{(sp)}$ - x, $\rho_{(t)}$ - *
Source: own work

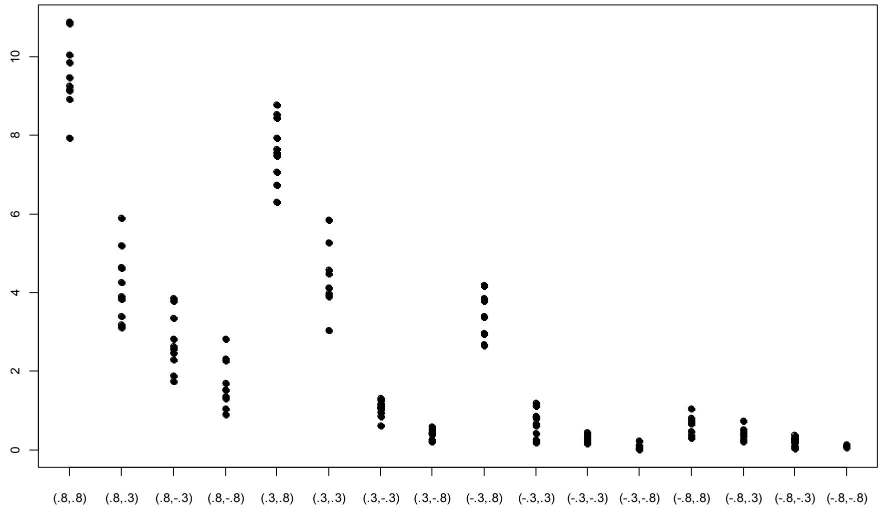

Figure 3. Increase of MSE of 10 domain total's predictors due to the estimation of model parameters (in %)
Source: own work

It is known (e.g. Datta, Lahiri (2000)) that the biases of REML estimators are of order o(D^{-1}) – in the simulation study, although the number of domains is small D=10, the biases (see Fig. 1) are small. In the Fig. 2 absolute values of RMSEs of the estimators are presented. Comparing them with real values ($\sigma_{\varepsilon}^{2}=1$, $\sigma_{u}^{2}=1$ and different values of $\rho_{(sp)}$ and $\rho_{(t)}$: 0,8; 0,3; -0,3 and -0,8) shows that the values are high.

Let us denote the predictor (8), where model parameters are known, as BLUP and the predictor (8), where model parameters are replaced by REML estimates, by EBLUP. In the Fig. 3 for each out of sixteen cases (defined by pairs of ($\rho_{(sp)},\rho_{(t)}$)) ten values of 100*(MSE(EBLUP)-MSE(BLUP))/MSE(BLUP) are presented for D=10 domains. The values of 100*(MSE(EBLUP)-MSE(BLUP))/MSE(BLUP) can be interpreted as the increase of the MSE (the decrease of accuracy) in % of the predictors of domain totals due to the estimation of model parameters. Although the MSEs of the estimators of model parameters are high (as presented in th Fig. 2) the increase of the MSE of domains' total predictors is small (as presented in the Fig. 3).

V.CONCLUSION

In the paper the problem of estimation of parameters of some longitudinal model is considered. The parameters are estimated using Restricted Maximum Likelihood Method by maximization of the log restricted likelihood using constrOptim R function. In the Monte Carlo simulation study values of the biases and RMSEs are computed. Although the RMSEs of the estimators are large, the influence of estimation of model parameters on the increase of the MSEs of domain totals' predictors is small.

REFERENCES

Chandra H., Salvati N., Chambers R. (2007), Small area estimation for spatially correlated populations – a comparison of direct and indirect model-based methods. *Statistics in Transition*, 8(2): 331–350.

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

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

Jiang, J. (1996), REML estimation: Asymptotic behavior and related topics, *The Annals of Statistics,* 24, 255–286.

Molina I., Salvati, N., Pratesi M. (2009), Bootstrap for estimating the MSE of the Spatial EBLUP. *Computational Statistics*, 24: 441–458.

Petrucci A., Salvati N. (2006), Small area estimation for spatial correlation in watershed erosion assessment. *J Agric Biol Environ Stat*, 11:169–182.

Petrucci A., Pratesi M., Salvati N. (2005), Geographic information in small area estimation: small area models and spatially correlated random area effects. *Statistics in Transition*, 7(3): 609–623.

Pratesi M., Salvati N. (2008), Small area estimation: the EBLUP estimator based on spatially correlated random area effects. *Stat Methods Appl*, 17: 113–141.

R Development Core Team (2011), A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna.

Salvati N., Pratesi M., Tzavidis N., Chambers R. (2009), Spatial M-quantile models for small area estimation. *Statistics in Transition*, 10(2), 251–261.

Thompson W.A., Jr. (1962), The problem of negative estimates of variance components, *Annals of Mathematical Statistics*, 33, 273–289.

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

Tomasz Żądło

O ESTYMACJI PARAMETRÓW PEWNEGO MODELU DLA DANYCH WIELOOKRESOWYCH

Rozważany jest problem modelowania profili wielookresowych zakładając, że populacja i przynależność elementów domen mogą zmieniać się w czasie. Proponowany model jest przypadkiem szczególnym ogólnego modelu liniowego i ogólnego mieszanego modelu liniowego. W modelu tym uwzględniono dwa wektory składników losowych spełniające odpowiednio założenia przestrzennego modelu autoregresyjnego i modelu autoregresyjnego rzędu pierwszego w czasie. W symulacji rozważano dokładność estymatorów parametrów modelu uzyskanych metodą największej wiarygodności z ograniczeniami.