The diagnostics of the mean squared error of the Eblup in small area estimation models

Pagliarella M.C. and Salvatore R.

Abstract In this paper, some issues related to the diagnostics of the mean squared error (MSE) of the Empirical Best Linear Unbiased Predictor (Eblup) in model-based small area estimation are considered. In particular, we develop special diagnostic tools that show the impact on the MSE of the Eblup caused by the deletion of areas in turn in the available data. The results are valid in general for the linear mixed model framework.

Key words: linear mixed model, empirical best linear unbiased predictor, mean squared error, diagnostics

1 The general linear mixed model

Some relevant aspects in the application of the small area estimation models can be assessed when we try to show the impact that some data have on the model estimation itself. Important features of the estimated small area models are connected with the evaluation on the data structure, concerning: a) their impact on the estimation of fixed and random effects, and on the covariance parameters estimates, then b), on the estimation of predicted values by the model, namely the survey parameters estimates, and c), on the estimation of the Mean Squared Error (MSE) of the Empirical Best Linear Unbiased Predictor (Eblup). The main feature of the paper is to highlight the impact of some subjects (areas, or cluster of areas in time-dependent data)

Maria Chiara Pagliarella
Università di Cassino e del Lazio Meridionale, e-mail: mc.pagliarella@unicas.it

Renato Salvatore
Università di Cassino e del Lazio Meridionale, e-mail: rsalvatore@unicas.it

We are grateful to Domingo Morales, Universidad Miguel Hernández de Elche (Spain), for his enlightening comments. We wish to thank the Italian National Institute of Statistics for providing the data.

on the MSE of the Eblup estimator in area-level models. The study is conducted as cluster (subject) deletion diagnostics analysis on every component of the MSE of the Eblup, in the case of the restricted maximum likelihood (Reml) estimation method. We give all the results in the following general linear model framework [2]:

$$y = X\beta + Zv + e, \tag{1}$$

with $y = col(y_i)$, $y_i = \widehat{\mu}_i$ $(i = 1,\dots,m, \sum n_i = n)$, being $\widehat{\mu}_i$ the vector (or a scalar, if $n_i = 1, \forall i$) of the direct estimators at the subject i. Further, $Z = diag(Z_1,\dots,Z_m)$, $v_1,\dots,v_m \overset{\text{iid}}{\sim} N(0,G)$, $e_i \overset{\text{ind}}{\sim} N(0,\Psi_i)$, with Ψ_i the covariance matrix (or a variance scalar, if $n_i = 1, \forall i$) of the sampling errors at the subject j. The model covariance is $V = ZDZ' + R$, with $D = I_m \otimes G$, and $R = \oplus_{i=1}^{m} \Psi_i$. Denoting by $\widehat{\theta}_R$ the vector of the covariance parameters of Reml estimates of the model 1, we consider as first step the MSE of the Eblup estimator $\widehat{\mu}^H = t(\widehat{\theta}_R)$ in the "naive approach" form [4]:

$$\begin{aligned}
MSE(\widehat{\mu}_i^H) &= MSE[t_i(\widehat{\theta}_R)] = g_{1i}(\widehat{\theta}_R) + g_{2i}(\widehat{\theta}_R),\\
t_i(\widehat{\theta}_R) &= p_i'\widehat{\beta} + m_i'\widehat{v}, && (2)\\
g_{1i}(\widehat{\theta}_R) &= m_i(G - GZ_i'V_i^{-1}Z_iG)m_i',\\
g_{2i}(\widehat{\theta}_R) &= (p_i - m_iG_iZ_i'V_i^{-1}X_i)M(p_i - m_iG_iZ_i'V_i^{-1}X_i)' = d_iMd_i', && (3)
\end{aligned}$$

being $M = (\sum X_i'V_i^{-1}X_i)^{-1}$, and p_i, m_i some vector or matrices that define the Eblup estimator 2. Consider now the component g_{1i}. Even in the case of known (estimated) $\widehat{\theta}_R$, we seemingly have an independence of the MSE estimation at subject i by the specific jth area (subject) deletion. However, the influence of its matrix Ψ_j of direct estimators sampling variances in the likelihood-based estimation can be relevant. Denoting by $g_{1i(j)}$ the first component for the MSE of the Eblup at the ith area, with the jth area deleted in the model 1, we have $g_{1i} - g_{1i(j)} = f(\widehat{\theta}_R - \widehat{\theta}_{R(j)})$, being $\widehat{\theta}_{R(j)}$ the restricted maximum likelihood (Reml) estimate of θ when the jth area is deleted. If we denote the joint residual log-likelihood function $l_R = \sum_{i\neq j} l_{Ri} + \delta l_{Rj}$ and the vector of the score equations $\sum_{i\neq j} S_{\theta i} + \delta S_{\theta j} = 0$, being S_θ the score equations for θ, and δ a control variable (scalar), the infinitesimal subject (area) deletion for $\widehat{\theta}_R$ is evaluated at:

$$\frac{d\widehat{\theta}_R(\delta)}{d\delta}|_{\delta=1} = -F_\theta^{-1}S_{\theta j} = -\left(\frac{\partial S_\theta}{\partial\theta}\right)^{-1}\frac{\partial S_\theta}{\partial\delta}|_{\delta=1} = -\left(\frac{\partial S_\theta}{\partial\theta}\right)^{-1}\frac{\partial l_{Ri}}{\partial\theta}. \tag{4}$$

Here F_θ is the θ-block of the Fisher information matrix, and F_θ^{-1} is the asymptotic covariance matrix of the estimates $\widehat{\theta}_R$. Further, we have the following relations, in terms of the Maximum likelihood (Ml) function l and of the Reml function l_R:

$$\begin{aligned}
\frac{\partial l_{Rj}}{\partial\theta} &= \frac{\partial l_j}{\partial\theta} + \frac{\partial\Delta_R}{\partial\theta} = -\frac{1}{2}\left(\frac{\partial vecV_j}{\partial\theta}\right)' vec(V_j^{-1} - V_j^{-1}r_jr_j'V_j^{-1}) + \frac{\partial\Delta_R}{\partial\theta} && (5)\\
\frac{\partial\Delta_R}{\partial\theta} &= \frac{1}{2}\left(\frac{\partial vecV}{\partial\theta}\right)' vec\left(P_{j,0} - P\right),
\end{aligned}$$

with $r_j = \widehat{\mu} - X_j\widehat{\beta}$, $P = V^{-1} - V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1} = V^{-1} - V^{-1}Q$, and Q the affine projection matrix using the distance induced by the matrix V^{-1}. The matrix $P_{j,0}$ is defined by the affine projection matrix $Q_{j,0}$, when we consider the $(n-1)$-dimensional projection subspace spanned by the columns of the matrix $X_{j,0}$. The last has the same elements of X, with a vector of zeroes at the row j. The magnitude of infinitesimal jth cluster (area) deletion for the vector of covariance parameters is consequently represented by the Hessian-normalized jth Ml score equation, plus the specific contribution due to the log-likelihood extra-term in the Reml function, respect to the Ml function. Now noting that we get for $\delta = 1$ by the log-likelihood the actual Reml estimate for θ, we can approximate the cluster (area) deletion diagnostics by a Taylor expansion of the differentiable function $\widehat{\theta}_R(\delta)$, as:

$$\widehat{\theta}_R(\delta) \approx \widehat{\theta}_R(1) + \frac{d\widehat{\theta}_R(\delta)}{d\delta}|_{\delta=1}(\delta - 1) = \widehat{\theta}_R(1) - F_\theta^{-1}S_{\theta j}(\delta - 1). \quad (6)$$

The function $\widehat{\theta}_R(\delta)$ at the point $\delta = 0$ is the Taylor approximation of the deletion diagnostics, $\widehat{\theta}_R(0)$, for the vector of the model covariance parameters at the jth area: $\widehat{\theta}_{R(j)} \approx \widehat{\theta}_R + F_\theta^{-1}S_{\theta j}$. We have, finally, the following deletion diagnostics for the first component of the MSE of the Eblup: $g_{1i(j)} - g_{1i} = g_{1i}(\widehat{\theta}_{R(j)}) - g_{1i}(\widehat{\theta}_R)$. Note that the relations 4–6, concerning the subject deletion diagnostics on the model covariance parameters, are valid in the general settings of the linear mixed model framework, when we deal with the Reml estimation method. In order to find the contribution to the MSE due to the variation of the estimator of the model fixed effects, noting that the deletion of the jth area yields:$g_{2i(j)} = d_i(\sum X_i'V_i^{-1}X_i - X_j'V_j^{-1}X_j)^{-1}d_i' = d_i(M - X_j'V_j^{-1}X_j)^{-1}d_i'$, and considering that we have $(M - X_j'V_j^{-1}X_j)^{-1} = M^{-1} + M^{-1}X_j'(V_j - X_jM^{-1}X_j')^{-1}X_jM^{-1}$, we come to $g_{2i(j)} - d_iMd_i' = g_{2i(j)} - g_{2i} = d_iM^{-1}X_j'(V_j - X_jM^{-1}X_j')^{-1}X_jM^{-1}d_i'$. The deletion of the jth area (subject) leads to the following useful relations:

$$\begin{aligned}
a)\; g_{2i(j)} - g_{2i} &= d_iM^{-1}X_j'V_j^{-1}(I_m - H_{1j})^{-1}X_jM^{-1}d_i' \\
H_{1j} &= X_jM^{-1}X_j'V_j^{-1} \\
b)\; g_{2i(j)} - g_{2i} &= d_i(\widehat{\beta} - \widehat{\beta}_{(j)})r_j'(r_jr_j')^{-1}X_jM^{-1}d_i' \\
\widehat{\beta} - \widehat{\beta}_{(j)} &= M^{-1}X_j'V_j^{-1}(I_m - H_{1j})^{-1}r_j \\
c)\; g_{2i(j)} - g_{2i} &= d_iM^{-1}X_j'V_j^{-1}H_{2j}{}^{-1}G_jV_j^{-1}X_jM^{-1}d_i' \\
H_{2j} &= Z_jG_jZ_j'V_j^{-1}(I_m - H_{1j}) = G_jV_j^{-1}(I_m - H_{1j}).
\end{aligned}$$

The matrix H_{1j} is the leverage matrix of the model fixed effects for the cluster j, $(\widehat{\beta} - \widehat{\beta}_{(j)})$ and r_j are, respectively, the cluster deletion diagnostics for the model general least squares estimate and the model residual at subject (area) j, H_{2j} is the leverage matrix associate to the random area (subject) effect. Combining the above results, we have the following MSE area (subject) deletion diagnostics:

$$\begin{aligned} MSE_{i(j)} - MSE_i &= (g_{1i(j)} - g_{1i}) + (g_{2i(j)} - g_{2i}) = \\ &= g_{1i}(\widehat{\theta}_{R(j)}) - g_{1i}(\widehat{\theta}_R) + (g_{2i(j)} - g_{2i}) \end{aligned}$$

We can extend the above deletion diagnostics of the MSE of the Eblup for the term $g_{3i(j)} - g_{3i}$ [4]. This diagnostics is related to the variability of the vector of the covariance parameters of Reml estimates $\widehat{\theta}_{R(j)}$, when we delete the jth area. We can easily utilize the above Taylor approximation (6), even we need an estimate of the asymptotic covariance matrix of $\widehat{\theta}_{R(j)}$.

2 Empirical study

We discuss an application of the MSE diagnostics in the context of a simple Fay-Herriot area-level model. The data are the official records from the Farm Structure Sample survey of the year 2007, collected by the Italian National Institute of Statistics. The aim is to analyze the impact of some administrative Italian provinces on the several components of the MSE of the Eblup, as highlighted in Fig.1. The dataset contains 103 observations. The target variable is the mean of standard gross profit, while the selected auxiliary variables are the mean of irrigable area and the number of the total working days.

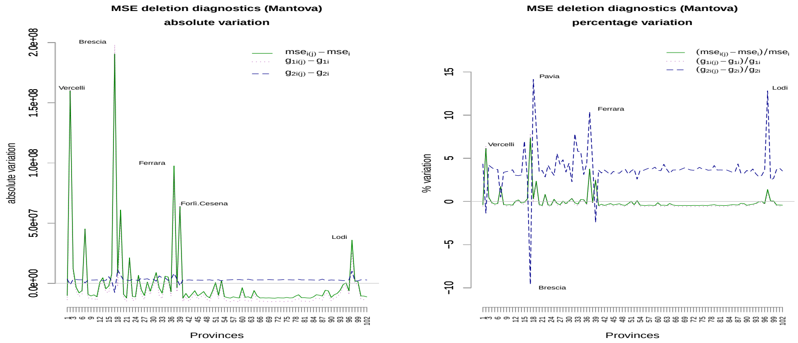

Fig.1 MSE deletion diagnostics at the Mantova province.

References

1. Demidenko, E., Stukel, T.A.: Influence analysis for linear mixed-effects models. Statistics in Medicine **24**, 893–909 (2005)
2. Demidenko, E.: Mixed Models, Theory and Applications. Wiley, New York, (2004)
3. Pregibon, D.: Logistic regression diagnostics. Annals of Statistics **4**, 705–724 (1981)
4. Rao, J.N.K.: Small Area Estimation. Wiley, New York, (2003)