Journal of Multivariate Analysis 106 (2012) 1–16

Contents lists available at SciVerse ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Parametric bootstrap methods for bias correction in linear mixed models

Tatsuya Kubokawa [a,*], Bui Nagashima [b]

[a] *Faculty of Economics, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan*

[b] *Graduate School of Economics, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan*

ARTICLE INFO

Article history:
Received 22 April 2011
Available online 16 December 2011

AMS 2010 subject classification:
62F40
62J07
62F12
62F15

Keywords:
Best linear unbiased predictor
Confidence interval
Empirical Bayes procedure
Fay–Herriot model
Second-order correction
Linear mixed model
Maximum likelihood estimator
Mean squared error
Nested error regression model
Parametric bootstrap
Restricted maximum likelihood estimator
Small area estimation

ABSTRACT

The empirical best linear unbiased predictor (EBLUP) in the linear mixed model (LMM) is useful for the small area estimation, and the estimation of the mean squared error (MSE) of EBLUP is important as a measure of uncertainty of EBLUP. To obtain a second-order unbiased estimator of the MSE, the second-order bias correction has been derived based on Taylor series expansions. However, this approach is hard to implement in complicated models with many unknown parameters like variance components, since we need to compute asymptotic bias, variance and covariance for estimators of unknown parameters as well as partial derivatives of some quantities. A similar difficulty occurs in the construction of confidence intervals based on EBLUP with second-order correction and in the derivation of second-order bias correction in the Akaike Information Criterion (AIC) and the conditional AIC. To avoid such difficulty in the derivation of second-order bias correction in these problems, the parametric bootstrap methods are suggested in this paper, and their second-order justifications are established. Finally, performances of the suggested procedures are numerically investigated in comparison with some existing procedures given in the literature.

1. Introduction

The linear mixed models (LMM) and the model-based estimates including empirical best linear unbiased predictor (EBLUP) or the empirical Bayes estimator (EB) have been recognized useful in small area estimation. The typical models used for the small area estimation are the Fay–Herriot model and the nested error regression model (NERM), and the usefulness of EBLUP is illustrated by Fay and Herriot [15] and Battese et al. [4]. For a good review and account on this topic, see [16,26,24].

When EBLUP is used to estimate a small area mean based on real data, it is important to estimate the mean squared error (MSE) since it can assess how much EBLUP is reliable. Asymptotically unbiased estimators of the MSE with the second-order bias correction have been derived based on the Taylor series expansion by Prasad and Rao [25], Datta and Lahiri [11], Datta et al. [12], Das et al. [8], Kubokawa [20] and others. A drawback of this method is that it is harder to compute the second-order bias, variance and covariance of estimators of more unknown parameters including variance components, and that it is troublesome to derive partial derivatives of some matrices with respect to unknown parameters. To avoid

* Corresponding author.
E-mail addresses: tatsuya@e.u-tokyo.ac.jp (T. Kubokawa), bui-bui.miu-miu@kde.biglobe.ne.jp (B. Nagashima).

doi:10.1016/j.jmva.2011.12.002

this difficulty, Butar and Lahiri [5] proposed the parametric bootstrap method, which is easy to implement, since we do not have to compute the second-order bias, variance and partial derivatives. For some recent results including nonparametric methods, see [22,17,7].

In the construction of a confidence interval based on EBLUP with the second-order accuracy, we are faced with a similar problem. Basu et al. [3], Datta et al. [9], Kubokawa [19] derived such confidence intervals using the Taylor series expansion. To avoid the difficulty in derivation of second-order moments, Chatterjee et al. [6], Hall and Maiti [18] proposed the confidence intervals using the parametric bootstrap method.

A similar difficulty occurs in evaluating the bias terms of *AIC* and conditional *AIC*. The Akaike Information Criterion (AIC) originated from [1,2] is recognized very useful for selecting models in general situations, and it is also useful for selecting variables in LMM. When unknown parameters in the model are estimated by the maximum likelihood estimator, the penalty term, which is a kind of bias, is known to be $2 \times p$ for dimension p of unknown parameters. When the unknown parameters in LMM are estimated by other estimators, however, Kubokawa [21] showed that the penalty term includes partial derivatives of the estimator and the covariance matrix. Concerning the conditional *AIC*, on the other hand, Liang et al. [23] Vaida and Blanchard [29] proposed the conditional *AIC* in LMM, but their derivations were limited to the cases that the parameters in LMM are partly known. Recently, Kubokawa [21] derived the second-order bias correction for the conditional *AIC*, but it is harder to compute in more complicated models.

In this paper, we treat the problems mentioned above, and provide useful procedures based on the parametric bootstrap methods to avoid the computational difficulties. In Section 2, we suggest the MSE estimator, the confidence interval, *AIC* and the conditional *AIC* using the parametric bootstrap. Concerning the MSE estimation, Butar and Lahiri [5] estimated the third term of the MSE, denoted by g_3, based on the parametric bootstrap, while in this paper, we consider to estimate the second-order approximation of g_3 using the parametric bootstrap method. A similar approach applies to the confidence interval, and we estimate the second-order correction term based on the parametric bootstrap method. This is different from the parametric bootstrap procedure suggested by Chatterjee et al. [6] who obtained two end-points of a confidence interval based on a distribution generated by the parametric bootstrap sampling. Simulation and empirical studies are given in Section 3. The proofs for the second-order justifications of the proposed procedures are given in the Appendix.

2. MSE estimation, confidence interval and *AIC* based on the parametric bootstrap method

2.1. Linear mixed model and the parametric bootstrap method

Consider the following general linear mixed model.

[1] *Model* 1. An $N \times 1$ observation vector $\boldsymbol{y}$ of the response variable has the model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{v} + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{X}$ and $\boldsymbol{Z}$ are $N \times p$ and $N \times M$ matrices, respectively, of the explanatory variables, $\boldsymbol{\beta}$ is a $p \times 1$ unknown vector of the regression coefficients, $\boldsymbol{v}$ is an $M \times 1$ vector of the random effects, and $\boldsymbol{\epsilon}$ is an $N \times 1$ vector of the random errors. Here, $\boldsymbol{v}$ and $\boldsymbol{\epsilon}$ are mutually independently distributed as $\boldsymbol{v} \sim \mathcal{N}_M(\boldsymbol{0}, \boldsymbol{G}(\boldsymbol{\theta}))$ and $\boldsymbol{\epsilon} \sim \mathcal{N}_N(\boldsymbol{0}, \boldsymbol{R}(\boldsymbol{\theta}))$, where $\boldsymbol{\theta} = (\theta_1, \dots, \theta_q)'$ is a q-dimensional vector of unknown parameters, and $\boldsymbol{G} = \boldsymbol{G}(\boldsymbol{\theta})$ and $\boldsymbol{R} = \boldsymbol{R}(\boldsymbol{\theta})$ are positive definite matrices. Then, $\boldsymbol{y}$ has a marginal distribution $\mathcal{N}_N(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$ for

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{R}(\boldsymbol{\theta}) + \boldsymbol{Z}\boldsymbol{G}(\boldsymbol{\theta})\boldsymbol{Z}'.$$

Throughout the paper, we often drop $(\boldsymbol{\theta})$ in $\boldsymbol{G}(\boldsymbol{\theta})$, $\boldsymbol{R}(\boldsymbol{\theta})$, $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ and others for notational convenience.

For simplicity, it is here assumed that $\boldsymbol{X}$ is of full rank. However, the results given in this paper can be extended to the case that $\boldsymbol{X}$ is not of full rank by modifying $\boldsymbol{X}$ and $\boldsymbol{\beta}$ as follows. Let r be a rank of $\boldsymbol{X}$, and suppose that $r < p$. Then, $\boldsymbol{X}$ can be written as

$$\boldsymbol{X} = \boldsymbol{P}\begin{pmatrix} \boldsymbol{E}_r & \boldsymbol{0} \\ \boldsymbol{0}' & \boldsymbol{0} \end{pmatrix}\boldsymbol{Q},$$

where $\boldsymbol{P}$ and $\boldsymbol{Q}$ are, respectively, $N \times N$ and $p \times p$ orthogonal matrices and $\boldsymbol{E}_r = \operatorname{diag}(\lambda_1, \dots, \lambda_r)$ for positive constants λ_i's. Let $\boldsymbol{P} = (\boldsymbol{P}_1, \boldsymbol{P}_2)$ and $\boldsymbol{Q}' = (\boldsymbol{Q}_1', \boldsymbol{Q}_2')$ for $N \times r$ matrix $\boldsymbol{P}_1$ and $r \times p$ matrix $\boldsymbol{Q}_1$. Thus, $\boldsymbol{X}\boldsymbol{\beta}$ can be rewritten as

$$\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{P}_1\boldsymbol{E}_r\boldsymbol{Q}_1\boldsymbol{\beta}.$$

Hence, all the results given in this paper still hold if we replace $(\boldsymbol{X}, \boldsymbol{\beta}, p)$ with $(\boldsymbol{P}_1\boldsymbol{E}_r, \boldsymbol{Q}_1\boldsymbol{\beta}, r)$.

The unknown parameters in Model 1 are $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. When $\boldsymbol{\theta}$ is known, the regression coefficients vector $\boldsymbol{\beta}$ is estimated by the generalized least squares estimator given by

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}) = (\boldsymbol{X}'\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\boldsymbol{y}.$$

The parameter $\boldsymbol{\theta}$ consists of variance components and others, and it is estimated by consistent estimator $\widehat{\boldsymbol{\theta}}$ based on $\boldsymbol{y}$ which can be constructed by maximum likelihood, restricted maximum likelihood and other methods. Then, $\boldsymbol{\beta}$ is estimated by $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\theta}})$.