British Journal of Mathematics & Computer Science
6(5): 370-380, 2015, Article no.BJMCS.2015.084
ISSN: 2231-0851

SCIENCEDOMAIN *international*
www.sciencedomain.org

SCIENCEDOMAIN

Linear Mixed Model in the Light of Future Data

Kunio Takezawa[1*]

[1] *Agroinformatics Division, Agricultural Research Center, National Agriculture and Food Research Organization Kannondai 3-1-1, Tsukuba, Ibaraki 305-8666, Japan.*

Article Information
DOI: 10.9734/BJMCS/2015/15514
Editor(s):
(1) Carlo Bianca, Department of Mathematical Sciences, Polytechnic University of Turin, Italy.
Reviewers:
(1) Yekti Widyaningsih, Department of Mathematics, University of Indonesia, Indonesia.
(2) Anonymous, Pakistan.
Complete Peer review History:
http://www.sciencedomain.org/review-history.php?iid=735&id=6&aid=7679

Original Research Article

Received: 01 December 2014
Accepted: 22 December 2014
Published: 09 January 2015

Abstract

The maximum likelihood and restricted (or residual) likelihood methods are common tools for estimating variances in linear mixed models. However, regression in the light of future data can yield different results. Investigations into the characteristics of this new variance are expected to promote the effective use of data in fields such as ecology and genetic statistics. Our numerical simulations show that the estimates of variances in the light of future data are substantially different from those given by the maximum likelihood and restricted (or residual) likelihood methods.

Keywords: Expected log-likelihood, linear mixed model, maximum likelihood estimator, optimization, third variance.

2010 Mathematics Subject Classification: 62E17; 62F10; 62J99

1 Introduction

Linear mixed models are regression analysis techniques that are based on extensive research. They have been successfully applied to various fields, and are regarded as a typical regression analysis technique. There is an extensive amount of literature and software available for this method (e.g.,

**Corresponding author: E-mail: nonpara@gmail.com;*

[1-6]). The regression coefficients in linear mixed models are treated as random variables from a normal distribution, so estimation methods for the variances of the normal distributions have an important role. The maximum likelihood (ML) and restricted (or residual) maximum likelihood (REML) methods are typically used for this purpose. The differences between the ML and REML methods are described in Section 6.10 of [7]. In short, the estimate of variance given by ML is an extension of maximum likelihood variance and that given by REML is an extension of unbiased variance. REML is usually used in practical applications of mixed model.

An alternative method is the "third variance", which was suggested in [8] and Section 5 of [9]. It can be used in combination with the maximum-likelihood and unbiased estimators of the variance. The third variance is based on "regression in the light of future data". The concept of "regression in the light of future data" is closely related to expected log-likelihood (Section 3.2 of [10]). That is, "regression in the light of future data" aims to maximize the expected log-likelihood given by a regression equation. The third variance is given by the simplest regression: estimation of variance when a nomal distribution is assumed. By introducing this method into variance estimates for linear mixed models we expect to yield better regression equations in terms of predictions, when compared to the ML or REML methods. In this paper, we investigate this claim using simple simulations.

In Section 2, we give the results of applying `lmer()` contained in the R (version 2.15.1) package "lme4 (version 1.0.4)", which produces simple linear mixed models using ML. The results are compared with those using a grid search optimization of the variance in terms of the log-likelihood. These procedures confirm that the ML results given by `lmer()` were the same as those given by maximizing the log-likelihood. The REML results given by `lmer()` were also confirmed to be accurate in a similar manner. Our R programs for these validations can be used as a basis for carrying out regression in the light of future data. Section 3 describes our numerical simulations, where we assumed that the constant multiplication of the variance given by ML or REML is optimal when estimating linear mixed models in the light of future data.

2 The "lme4" Packages with ML and REML

The work in this paper considered the simplest regression equation of the linear mixed model. That is,

$$\mathbf{y} = \boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}. \tag{2.1}$$

This equation is derived by eliminating the linear part of the typical linear mixed model (e.g., page 98 of [2]), which is an extension of a simple regression equation. This simple regression equation is adopted because it makes the number of parameters to be estimated smaller. If the number of parameters is large, grid search for the best parameters needs long period of time.

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ \vdots \\ y_n \end{pmatrix}, \qquad \mathbf{Z} = \begin{pmatrix} \mathbf{1}_{s\times 1} & \mathbf{0}_{s\times 1} & \cdots & \mathbf{0}_{s\times 1} \\ \mathbf{0}_{s\times 1} & \mathbf{1}_{s\times 1} & \cdots & \mathbf{0}_{s\times 1} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{s\times 1} & \mathbf{0}_{s\times 1} & \cdots & \mathbf{1}_{s\times 1} \end{pmatrix}, \tag{2.2}$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta \\ \beta \\ \vdots \\ \vdots \\ \beta \end{pmatrix}, \qquad \mathbf{u} = \begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_m \end{pmatrix}, \qquad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \vdots \\ \epsilon_n \end{pmatrix}, \tag{2.3}$$

where $\mathbf{y}$ is the target variablefs data. $\mathbf{Z}$ contains $\mathbf{1}_{s\times 1}$ (a column vector comprising s elements of 1) and $\mathbf{0}_{s\times 1}$ (a column vector comprising s elements of 0). $\boldsymbol{\beta}$ is a column vector comprising n elements of β. $\{U_i\}(1 \leq i \leq m)$ are elements of $\mathbf{u}$ and are realizations of $N(0, \sigma_U^2)$ (a normal distribution with mean 0 and variance σ_U^2). n is the number of data ($n = m \cdot s$).

Next, we set

$$\mathbf{G} = \sigma_U^2 \mathbf{I}_m, \qquad \mathbf{R} = \sigma_\epsilon^2 \mathbf{I}_n, \qquad \mathbf{V} \equiv cor(\mathbf{y}) = \mathbf{Z}\mathbf{G}\mathbf{Z}^t + \mathbf{R}, \tag{2.4}$$

where $\mathbf{I}_m$ is the $m \times m$ identity matrix, and $\mathbf{I}_n$ is the $n \times n$ identity matrix. The ML estimate of $\mathbf{V}$ is based on the model:

$$\mathbf{y} \sim N(\boldsymbol{\beta}, \mathbf{V}) \tag{2.5}$$

Then, the log-likelihood of Eq. (2.1) in the light of $\mathbf{y}$ is (see, for example, page 101 of [2])

$$l_P(\mathbf{V}) = -0.5\Big(\log(|\mathbf{V}|) + (\mathbf{y} - \boldsymbol{\beta})^t \mathbf{V}^{-1}(\mathbf{y} - \boldsymbol{\beta}) + n\log(2\pi)\Big). \tag{2.6}$$

The ML method maximizes this value for estimating β, σ_U^2, and σ_ϵ^2. This leads to the estimate of β (see, for example, page 163 of [6]),

$$\hat{\beta} = \frac{1}{n}\sum_{i=1}^{n} y_i. \tag{2.7}$$

Therefore, the value given by substituting Eq. (2.7) into Eq. (2.6) is maximized and used to derive σ_U^2 and σ_ϵ^2. We also use Eq. (2.7) when applying the REML method, because REML does not estimate $\hat{\beta}$ (see page 178 of [6]).

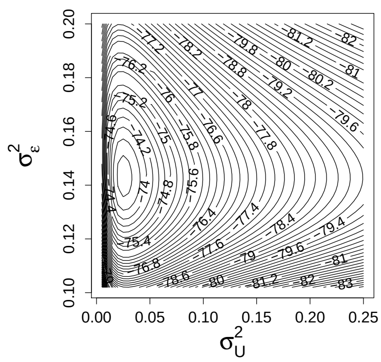

Figure 1: Log-likelihoods ($l_P(\mathbf{V})$(Eq.(2.6))) using various values of σ_U^2 and σ_ϵ^2. The log-likelihood in the area where σ_U^2 is small rises sharply.

Then, we compared the ML results given by `lmer()` (contained in the package "lme4 (version 1.0.4)") with those using the log-likelihood on the grid-point variance values.

We first produced regression equations in the form of Eq. (2.1). We set $\beta = 9.9$, $s = 15$, and $m = 10$ in Eq. (2.1) and generated simulation data. This results in $n = 150$. Furthermore, we used realizations of $N(0, 0.2^2)$ (a normal distribution with mean 0 and variance 0.2^2) as $\{U_i\}$, and realizations of $N(0, 0.4^2)$ (a normal distribution with mean 0 and variance 0.4^2) as $\{\epsilon_{ij}\}$. We obtained $\hat{\beta}$ using Eq. (2.7). The resulting value was substituted into Eq. (2.6). The value of σ_U^2 was set to one of 50 values ($\{0.005, 0.010, 0.015, \ldots, 0.25\}$). The value of σ_ϵ^2 was set to one of 100 values ($\{0.102, 0.104, 0.106, \ldots, 0.3\}$). Then, we calculated the value of Eq. (2.6) using one of the

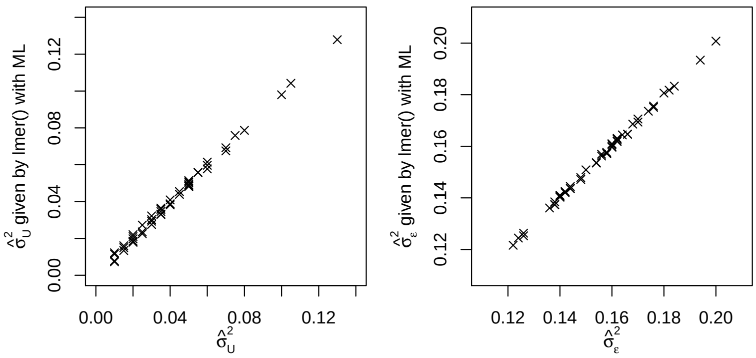

Figure 2: The horizontal axis represents σ_U^2 estimated using the grid-point values of $l_P(\mathbf{V})$(Eq. (2.6)). The vertical axis represents σ_U^2 that resulted from `lmer()`(left). The horizontal axis represents σ_ϵ^2 estimated using the grid-point values of $l_P(\mathbf{V})$(Eq. (2.6)). The vertical axis represents σ_ϵ^2 that resulted from `lmer()`(right).

grid-points derived from these values. The results are shown in Fig. 1. We compared the optimum values of σ_U^2 and σ_ϵ^2 determined using this procedure with those given by `lmer()` with the ML setting. This simulation was repeated 50 times using various initial values for the pseudo-random numbers. The results are shown in Fig. 2. The two sets of estimates are very similar. Hence, `lmer()` with the ML setting provides very accurate estimates using the ML method. We modified this R program for calculating the log-likelihood in the light of future data to create the program used in the next section.

Next, we assessed the validity of the results given by `lmer()` with the REML setting. Most of the settings were the same as in Fig. 2, although we used the following definition of $l_R(\mathbf{V})$ (see, for example, page 101 of [2]) instead of $l_P(\mathbf{V})$(Eq. (2.6)) to carry out the numerical simulations.

$$l_R(\mathbf{V}) = -0.5\bigg(\text{log}(|\mathbf{V}|) + (\mathbf{y} - \boldsymbol{\beta})^t\mathbf{V}^{-1}(\mathbf{y} - \boldsymbol{\beta}) + n\text{log}(2\pi) + \text{log}((\mathbf{1}_{n\times 1})^t\mathbf{V}^{-1}\mathbf{1}_{n\times 1})[RC4]\bigg), \quad (2.8)$$

where $\mathbf{1}_{n\times 1}$ is a column vector comprising n elements of 1. The results are shown in Fig. 3, and confirm that `lmer()` with the REML setting produces very accurate estimates.

3 Linear Mixed Model Given by Regression in the Light of Future Data

We define the future data as

$$\mathbf{y}^* = \begin{pmatrix} y_1^* \\ y_2^* \\ \vdots \\ \vdots \\ y_n^* \end{pmatrix}. \quad (3.1)$$

The log-likelihood of the regression equation derived by maximizing Eq. (2.6) in the light of $\mathbf{y}^*$ is

$$l_P^*(\hat{\mathbf{V}}) = -0.5\bigg(\text{log}(|\hat{\mathbf{V}}|) + (\mathbf{y}^* - \hat{\boldsymbol{\beta}})^t\hat{\mathbf{V}}^{-1}(\mathbf{y}^* - \hat{\boldsymbol{\beta}}) + n\text{log}(2\pi)\bigg). \quad (3.2)$$

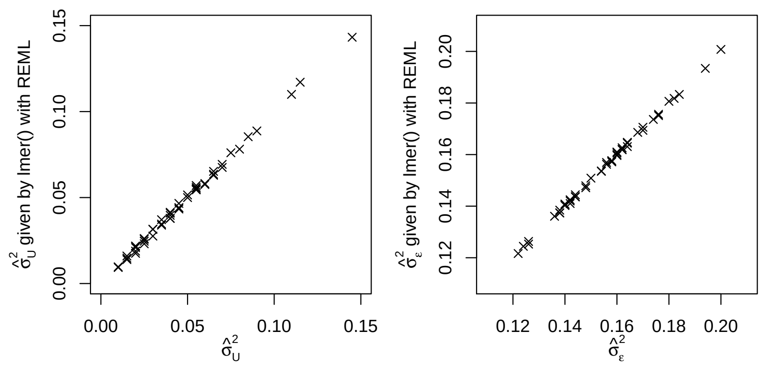

Figure 3: The horizontal axis represents σ_U^2 estimated by the grid-point values of $l_P(\mathbf{V})$(Eq. (2.8)). The vertical axis represents σ_U^2 that resulted from `lmer()`(left). The horizontal axis represents σ_ϵ^2 estimated by the grid-point values of $l_P(\mathbf{V})$(Eq.(2.8)). The vertical axis represents σ_ϵ^2 that resulted from `lmer()`(right).

σ_U^2 and σ_ϵ^2 that maximize the log-likelihood in the light of future data are $\alpha_U\hat{\sigma}_U^2$ and $\alpha_\epsilon\hat{\sigma}_\epsilon^2$. $\hat{\sigma}_U^2$ and $\hat{\sigma}_\epsilon^2$ are estimated using the available data and either REML or ML. Then, Eq. (2.4) is replaced with

$$\hat{\mathbf{G}} = \alpha_U\hat{\sigma}_U^2\mathbf{I}_m, \qquad \hat{\mathbf{R}} = \alpha_\epsilon\hat{\sigma}_\epsilon^2\mathbf{I}_n, \qquad \hat{\mathbf{V}} = \mathbf{Z}\hat{\mathbf{G}}\mathbf{Z}^t + \hat{\mathbf{R}}. \tag{3.3}$$

First, we set $\beta = 9.9$, $s = 15$, and $m = 10$. $\{U_i\}$ are realizations of $N(0, 0.2^2)$ (a normal distribution with mean 0 and variance 0.2^2), and $\{\epsilon_{ij}\}$ are realizations of $N(0, 0.4^2)$ (a normal distribution with mean 0 and variance 0.4^2). We used `lmer()`with the REML setting to estimate $\hat{\beta}$, $\hat{\sigma}_U^2$, and $\hat{\sigma}_\epsilon^2$. Using these three estimates, we derived $l_P^*(\hat{\mathbf{V}})$ (Eq.(3.2)). $\hat{\alpha}_U$ was set to be one of $\{1, 1.1, 1.2, \ldots, 1.9\}$, and $\hat{\alpha}_\epsilon$ was set to be one of $\{1, 1.01, 1.02, \ldots, 1.09\}$. We used the grid-points constructed by these two values for this simulation. $\mathbf{y}^*$ was obtained using the same conditions as for $\mathbf{y}$, although we used different initial values for the pseudo-random numbers. These different initial values produced 20 sets of $\mathbf{y}^*$ to derive $l_P^*(\hat{\mathbf{V}})$. We took the average of these 20 values. This numerical simulation was repeated 100 times using different initial values for the pseudo-random numbers, and we averaged the 100 mean values for $l_P^*(\hat{\mathbf{V}})$. The results are shown in Fig. 4. When the optimal values of $(\alpha_U, \alpha_\epsilon)$ were $(\hat{\alpha}_U, \hat{\alpha}_\epsilon)$, the means were $(\hat{\alpha}_U, \hat{\alpha}_\epsilon) = (1.325, 1.03)$. That is, $\hat{\alpha}_U\hat{\sigma}_U^2$ was substantially larger than the estimates given by the REML method. Although this tendency was not very apparent in $\hat{\sigma}_\epsilon^2$, $\hat{\alpha}_\epsilon$ was larger than 1.

Figure 5 shows the estimation results for $\hat{\beta}$, $\hat{\sigma}_U^2$, and $\hat{\sigma}_\epsilon^2$. We used `lmer()` with the ML setting, although the other conditions were the same as in Fig. 4. The mean values were $(\hat{\alpha}_U, \hat{\alpha}_\epsilon) = (1.525, 1.035)$. That is, $\hat{\alpha}_U\hat{\sigma}_U^2$ was substantially larger than that given by the ML method. Because $\hat{\sigma}_U^2$ calculated by the ML method was smaller than that given by the REML method, in this case, $\hat{\alpha}_U$ was larger than that given by the REML method. Furthermore, $\hat{\alpha}_\epsilon$ was larger than 1.

When $s = 10$ and $m = 5$, and the other conditions were the same as Fig. 4, we obtained the results shown in Fig. 6. The value of $\hat{\alpha}_U$ was set to one of 10 values ($\{1, 1.1, 1.2, \ldots, 1.9\}$), and $\hat{\alpha}_\epsilon$ was set to one of 10 values ($\{1, 1.05, 1.1, \ldots, 1.45\}$). We used one of grid-point values given by these values for this calculation. The means were $(\hat{\alpha}_U, \hat{\alpha}_\epsilon) = (1.4, 1.1125)$. As in the previous case, $\hat{\alpha}_U\hat{\sigma}_U^2$ was substantially larger than that given by the REML method. Furthermore, $\hat{\alpha}_\epsilon$ was larger than 1. Figure 7 shows the results of the numerical simulations with the ML setting instead of REML, although the other conditions are the same as in Fig. 5. The value of $\hat{\alpha}_U$ was set to one of 10 values ($\{1, 1.2, 1.4, \ldots, 2.8\}$), and $\hat{\alpha}_\epsilon$ was set to one of 10 values ($\{1, 1.05, 1.1, \ldots, 1.45\}$). We used one of

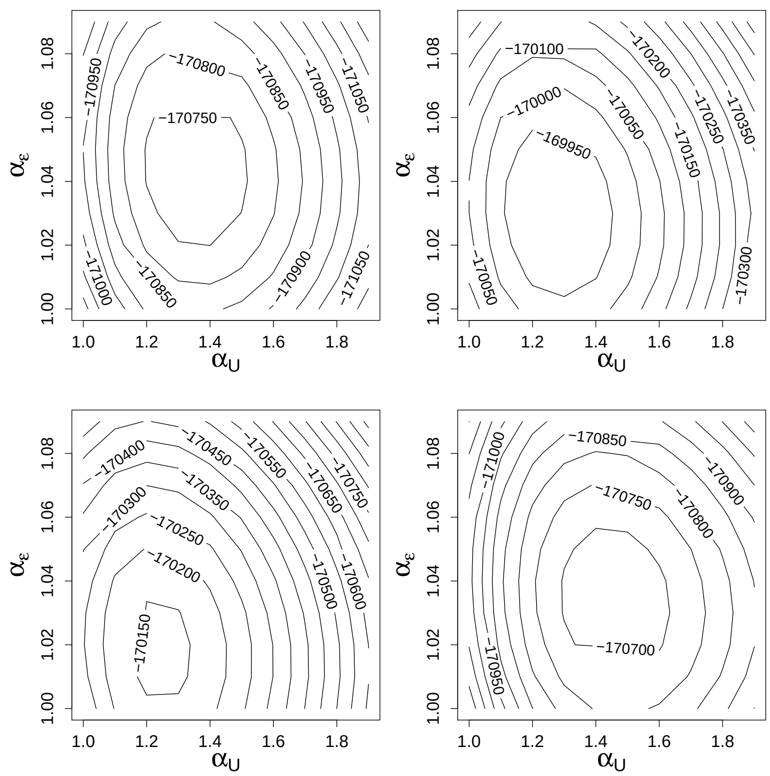

Figure 4: We repeated the numerical simulations four times using various initial values for the pseudo-random numbers. The mean values of the 4 sets were $(\hat{\alpha}_U, \hat{\alpha}_\epsilon) = (1.325, 1.03)$.

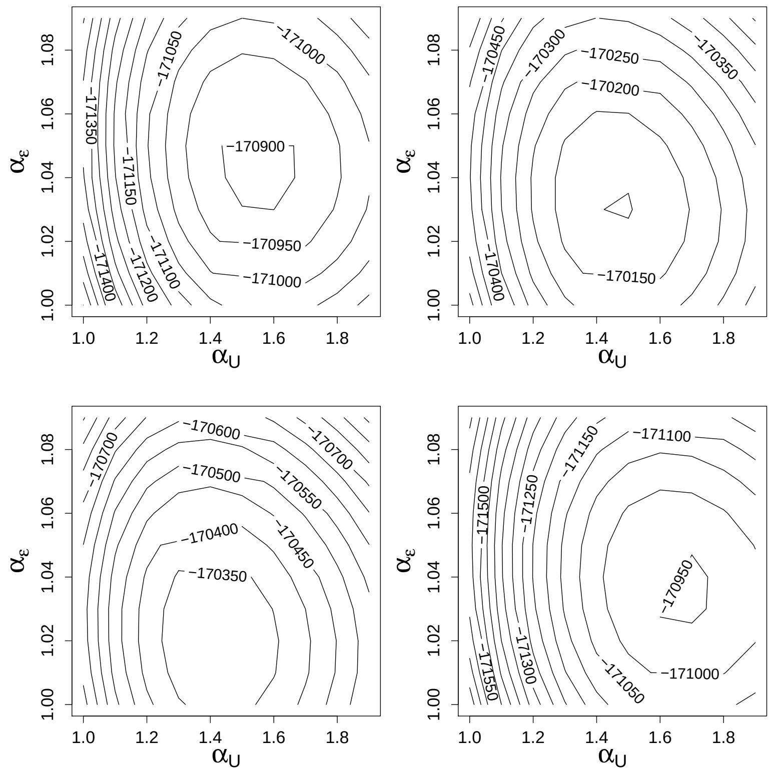

Figure 5: Mean of $l_P^*(\hat{\mathbf{V}})$ given by $\hat{\sigma}_U^2$ and $\hat{\sigma}_\epsilon^2$. We used the ML method to obtain these variances. In this experiment, $\beta = 9.9$, $s = 15$, and $m = 10$. The numerical simulations were repeated four times using various initial values for the pseudo-random numbers. The mean values were $(\hat{\alpha}_U, \hat{\alpha}_\epsilon) = (1.525, 1.035)$.

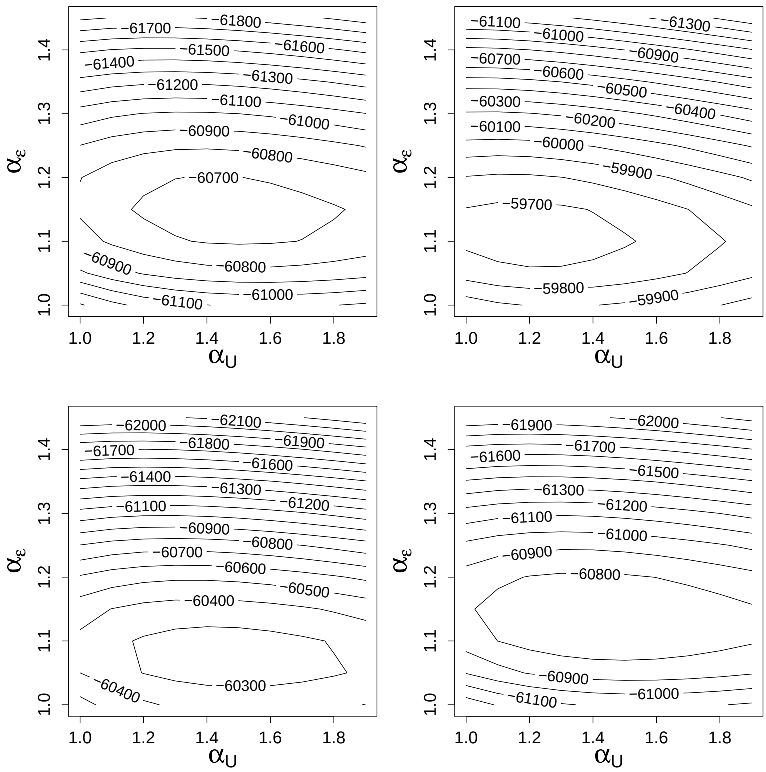

Figure 6: Mean of $l_P^*(\hat{\mathbf{V}})$ given by $\hat{\sigma}_U^2$ and $\hat{\sigma}_\epsilon^2$. We used the REML method to obtain these variances. Here, $\beta = 9.9$, $s = 15$, and $m = 10$. We repeated the numerical simulations four times using various initial values for the pseudo-random numbers. The mean values were $(\hat{\alpha}_U, \hat{\alpha}_\epsilon) = (1.4, 1.1125)$.

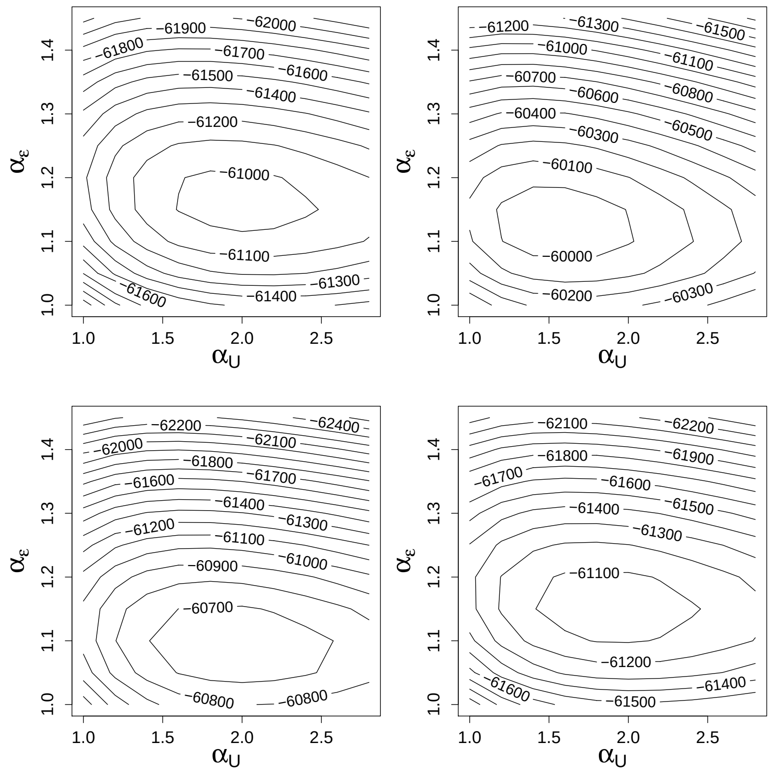

Figure 7: Mean of $l_P^*(\hat{\mathbf{V}})$ given by $\hat{\sigma}_U^2$ and $\hat{\sigma}_\epsilon^2$. We used the ML method to obtain these variances. Here, $\beta = 9.9$, $s = 15$, and $m = 10$. The numerical simulations were repeated four times using various initial values for the pseudo-random numbers. The mean values were $(\hat{\alpha}_U, \hat{\alpha}_\epsilon) = (1.85, 1.125)$.

the grid-points given by these values for this calculation. The means were $(\hat{\alpha}_U, \hat{\alpha}_\epsilon) = (1.85, 1.125)$. $\hat{\alpha}_U$ was larger than that given by the REML method. Furthermore, $\hat{\alpha}_\epsilon$ was larger than 1.

4 Conclusions

Our numerical simulations demonstrate that when $\hat{\sigma}^2_U$ is estimated in the light of future data, the estimate should be substantially larger than that given by the REML or ML methods. That is, a linear mixed model estimate requires the "third variance" concept (Takezawa (2012); Section 5 of Takezawa (2013)). Moreover, the difference between conventional variances and the variance that considers future data should not be ignored when using linear mixed models in practical applications. Values of 5 or 10 for m are typical when applying linear mixed models to fields such as ecology, genetic statistics, and animal breeding. In these applications, m stands for the number of cultivars or the number of replications in an experiment. This indicates that the estimates derived using the REML or ML methods are not valid, because the values of $\hat{\alpha}_U$ obtained here are substantially larger than 1. Therefore, the values of $\hat{\sigma}^2_U$ calculated using the REML or ML methods (and the estimates it gives for each cultivar or replication) are considerably different from those obtained by maximizing the log-likelihood in the light of future data. This suggests that existing linear mixed models have not completely exploited the information contained in available data.

However, the results obtained here are based on a limited number of numerical simulations. Most applications of linear mixed models are more complex than Eq.(2.1), and often include more linear term(s). Moreover, generalized linear mixed models are commonly used (in which errors follow a distribution other than the normal distribution). Hence, a broad range of simulations and analytical research should be carried out, so that we can fully understand the details of this problem and establish reliable methods for estimation in the light of future data.

Acknowledgements

The author is very grateful to the referees for carefully reading the paper and for their comments and suggestions which have improved the paper.

Competing Interests

The author declares that no competing interests exist.

References

[1] Pinheiro J, Bates D. Mixed-effects models in S and S-PLUS. Springer;2000, 2009.

[2] Ruppert D, Wand MP, Carroll RJ. Semiparametric regression. Cambridge University Press;2003.

[3] Faraway JJ. Extending the Linear Model with R: Generalized linear, mixed effects and nonparametric regression models. Chapman and Hall/CRC;2005.

[4] Wood S. Generalized additive models: An Introduction with R. Chapman and Hall/CRC; 2006.

[5] Rupport D, Ward MP, Carrol RJ. Semiparametric regression during 2003-2007. Electronic Journa of Statistics. 2009;3:1193-1256.

[6] Zuur AF, Ieno EN, Walker NJ, Saveliev AA, Smith GM. Mixed effects models and extensions in ecology with R. Springer;2009.

[7] McCulloch CE, Shayle R, Searle SR. Generalized, linear, and mixed models. Wiley; 2001.

[8] Takezawa K. A Revision of AIC for normal error models. Open Journal of Statistics. 2012;2(3):309-312.

[9] Takezawa K. Learning regression analysis by simulation. Springer; 2013.

[10] Konishi S, Kitagawa G. Information criteria and statistical modeling. Springer; 2008.

Peer-review history:
The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)
www.sciencedomain.org/review-history.php?iid=735&id=6&aid=7679