CIRJE-F-957

Box-Cox Transformed Linear Mixed Models for Positive-Valued and Clustered Data

Shonosuke Sugasawa
Graduate School of Economics, The University of Tokyo

Tatsuya Kubokawa
The University of Tokyo

February 2015

Box-Cox Transformed Linear Mixed Models for Positive-Valued and Clustered Data

Shonosuke Sugasawa* and Tatsuya Kubokawa†

Abstract

The Box-Cox transformation is applied to linear mixed models for analyzing positive and clustered data. The problem is that the maximum likelihood estimator of the transformation parameter is not consistent. To fix it, we suggest a simple and consistent estimator for the transformation parameter based on the moment method. The consistent estimator is used to construct consistent estimators of the parameters involved in the model and to provide an empirical predictor of a linear combination of both fixed and random effects. Second-order accurate prediction intervals for measuring uncertainty of the predictor are derived. Finally, the performance of the proposed procedure is investigated through simulation and empirical studies.

Key words and phrases: Box-Cox transformation, consistency, linear mixed model, parametric bootstrap, prediction intervals, second-order approximation, small-area estimation, transformation parameter, variance component.

1 Introduction

The linear mixed models (LMM) and the empirical best linear unbiased prediction (EBLUP) have been studied and extensively used in literature from theoretical and practical aspects. For example, see Baltagi (2013) for panel data analysis, Verbeke and Molenberghs (2009) for longitudinal data analysis and Rao (2003) for small area estimation. In particular, the linear mixed models are useful for analyzing clustered or grouped data, since they combine information from different sources to achieve satisfactory inference. For a good review on the linear mixed models and the related topics, see Jiang and Lahiri (2006) and McCulloch and Searle (2001).

In many surveys, the data are positive, such as income, revenue, harvest yield or production. The distributions are positively skewed, and need suitable transformations for normality to hold. The log-transformation is a standard method, but not always appropriate. A conventional alternative is the Box-Cox transformation (Box and Cox, 1962), described as

$$h(x,\lambda)=\begin{cases}\lambda^{-1}(x^{\lambda}-1), & \lambda\neq 0\\ \log x, & \lambda=0\end{cases} \tag{1}$$

*Graduate School of Economics, University of Tokyo, E-Mail: shonosuke622@gmail.com
†Faculty of Economics, University of Tokyo

for positive x. The transformation parameter λ enables us to transform positive data flexibly to fit normal models. However, the crucial issue is that the maximum likelihood (ML) estimator of λ is not consistent. This comes from the fact that the range of the Box-Cox transformation is not whole real line except for $\lambda = 0$ (log-transformation), which is known as the truncation problem. The inconsistency of the ML estimator influences the inference. Concerning the prediction problem in the linear mixed model with the Box-Cox transformation, we can construct the EBLUP by replacing the model parameters with their estimators in the best linear unbiased predictor (BLUP). Due to the inconsistency, the EBLUP based on the ML estimator does not converge to the BLUP. As investigated numerically in Section 4, this causes the large estimation errors. Concerning the estimation of λ, Gurka, Edwards, Muller and Kupper (2006) considered the likelihood inference for λ, and Lee, Lin, Lee and Hsu (2005) addressed the Bayesian inference for λ in the linear mixed models with time series dependence. Although both papers provided good methods for the inference, they did not investigate theoretical properties such as consistency. As the related transformation, Yang (2006) proposed the dual power transformation $h^{\mathrm{DPT}}(x, \lambda) = (x^{\lambda} - x^{-\lambda})/2\lambda$ for $\lambda > 0$; $= \log x$ for $\lambda = 0$ and established consistency of the ML estimator of λ. Sugasawa and Kubokawa (2014 a, b) used this transformation in the linear mixed models and provided quite reasonable analysis from theoretical and practical points of view. Since the dual power transformation is not so simple as the Box-Cox transformation, it might be harder to figure out how the data are transformed with the dual power transformation.

An advantage of the Box-Cox transformation is that we can catch up how the data are transformed, since it is almost the same as a power transformation. For example, the cases of $\lambda = 1$ and $\lambda = 0.5$ correspond to a linear and a squared root transformations, respectively. However, the inference based on the ML estimator of λ is influenced by a disadvantage of the inconsistency. Thus, in this paper, we suggest a new method for estimating the transformation parameter λ in the Box-Cox transformation. In the framework of the linear mixed models, the estimator is obtained as the solution of the equation

$$\sum_{i=1}^{m} \sum_{j=1}^{n_i} \left\{ h(y_{ij}, \lambda) - \boldsymbol{x}_{ij}' \widehat{\boldsymbol{\beta}}(\lambda) \right\}^3 = 0,$$

where the notations are explained in Section 2. This estimating equation is based on the moment method, since $h(y_{ij}, \lambda) - \boldsymbol{x}_{ij}'\boldsymbol{\beta}$ has a normal distribution and the third moment is zero. This method provides the consistency as established in Section 2 as well as it has a simpler formula than the ML method.

In Section 2, we describe the linear mixed model where the observed response data are transformed by the Box-Cox transformation. We call here this model the Box-Cox transformed linear mixed model (BC-LMM). The model includes the unknown parameters such as regression coefficients, variance components and the transformation parameter. These model parameters are estimated consistently based on the above moment estimator of λ.

The primary purpose of the linear mixed models is the prediction of a linear combination, denoted by μ, of fixed effects and random effects. In the Box-Cox transformed linear mixed models, we derive the best linear unbiased predictor (BLUP) and provide the empirical best linear unbiased predictor (EBLUP) $\widehat{\mu}^{\mathrm{EBLUP}}$ by plugging in the consistent estimators of the model parameters. However, an interesting quantity of prediction is the inversely transformed value of μ given by $h^{-1}(\mu, \lambda)$ and we can suggest the corresponding predictor $h^{-1}(\widehat{\mu}^{\mathrm{EBLUP}}, \widehat{\lambda})$, which is called the Box-Cox transformed EBLUP (BC-TEBLUP). In Section 3, we obtain a prediction interval of $h^{-1}(\mu, \lambda)$ based on the BC-TEBLUP with second-order accuracy $O(m^{-1})$ for the nominal confidence coefficient, where m is the number of clusters. This interval can be constructed by using the parametric bootstrap method given in Chatterjee, Lahiri and Lee (2008).

In Section 4, finite sample performances of the suggested procedures are investigated by simulation. In estimation of the model parameters, it is shown that the estimators based on the moment estimator of λ have much smaller mean squared errors than the corresponding estimators based on the ML estimator of λ. Concerning the prediction intervals based on the parametric bootstrap, its coverage probability can be shown to be close to the nominal confidence coefficient.

In Section 4, we also provide two illustrative examples: the rat growth data given in Gelfand, Hills, Racine-Poon and Smith (1990) and the posted land price data treated in Kawakubo and Kubokawa (2014). The Box-Cox transformed random coefficients model and the Box-Cox transformed nested error regression model are applied to the former and latter data, respectively. The concluding remarks are given in Section 5. Finally, technical proofs are given in Appendix.

2 Box-Cox Transformed Linear Mixed Models

2.1 Model settings

Consider the two-stage cluster sampling, namely m clusters are randomly selected, and positive data are randomly sampled from each selected cluster.

Let $y_{i1}, \ldots, y_{in_i}$ be positive observations sampled from the i-th cluster for $i = 1, \ldots, m$. For the Box-Cox transformation $h(x, \lambda)$ defined in (1), we denote $(h(y_{i1}, \lambda), \ldots, h(y_{in_i}, \lambda))'$ by $h(\boldsymbol{y}_i, \lambda)$ for $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})'$. The Box-Cox transformed linear mixed model is described as

$$h(\boldsymbol{y}_i, \lambda) = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, m \tag{2}$$

where $\boldsymbol{X}_i$ is an $n_i \times p$ known matrix, $\boldsymbol{Z}$ is an $n_i \times q$ known matrix and $\boldsymbol{\beta}$ is a p-variate unknown vector of regression coefficients. Here, $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m$ and $\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_m$ are mutually independently distributed as $\boldsymbol{b}_i \sim \mathcal{N}_q(\boldsymbol{0}, \boldsymbol{R})$ and $\boldsymbol{\epsilon}_i \sim \mathcal{N}_{n_i}(\boldsymbol{0}, \boldsymbol{G}_i)$. We further assume that $q \times q$ matrix $\boldsymbol{R}$ and $n_i \times n_i$ matrix $\boldsymbol{G}_i$ are known up to k unknown parameters, namely $\boldsymbol{R} = \boldsymbol{R}(\boldsymbol{\psi})$ and $\boldsymbol{G}_i = \boldsymbol{G}_i(\boldsymbol{\psi})$ for unknown parameters $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_k)'$. Thus the unknown model parameters are $\boldsymbol{\theta} = (\lambda, \boldsymbol{\beta}', \boldsymbol{\psi}')'$.

In the model (2), the first term $\boldsymbol{X}_i\boldsymbol{\beta}$ is fixed effects and the second term $\boldsymbol{Z}_i\boldsymbol{b}_i$ is random effects. Note that $\boldsymbol{\epsilon}_i$ is noise and $\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i$ is the signal component of interest of the observed data. The covariance matrix of the transformed observations $h(\boldsymbol{y}_i, \lambda)$ is given by

$$\boldsymbol{\Sigma}_i(\boldsymbol{\psi}) = \boldsymbol{Z}_i\boldsymbol{R}(\boldsymbol{\psi})\boldsymbol{Z}_i' + \boldsymbol{G}_i, \quad i = 1, \ldots, m.$$

The BC-LMM described in (2) includes various applicable models. The transformed nested error regression model and the transformed random coefficient model are two typical examples used in the framework of small area estimation. These models will be used in simulation and empirical studies given in Section 4.

1. Transformed nested error regression model.
The nested error regression model (NERM) proposed by Battese et al.(1988) is extensively used for small area estimation. The model (2) enables us to take the Box-Cox transformation in the NERM, namely Box-Cox transformed nested error regression model (BC-NERM) described as

$$h(y_{ij},\lambda)=\boldsymbol{x}_{ij}'\boldsymbol{\beta}+v_i+\varepsilon_{ij},\qquad j=1,\dots,n_i,\quad i=1,\dots,m, \tag{3}$$

where v_i's and ε_{ij}'s are random area effects and error components respectively, and mutually independently distributed $v_i\sim\mathcal{N}(\mathbf{0},\sigma_1^2)$ and $\varepsilon_{ij}\sim\mathcal{N}(\mathbf{0},\sigma_2^2)$. The variables and parameters in the form (2) correspond to $\boldsymbol{\psi}=(\sigma_1,\sigma_2)$, $\boldsymbol{Z}_i=\mathbf{1}_{n_i}$, $\boldsymbol{b}_i=v_i$, $\boldsymbol{R}(\boldsymbol{\psi})=\sigma_1$ and $\boldsymbol{G}_i=\sigma_2^2\boldsymbol{I}_{n_i}$. In empirical study in Section 4.4, we apply this model to the posted land price (PLP) data in Japan.

2. Transformed random coefficient model.
A random coefficient model is useful for analyzing growth data. Since such growth data take positive values, the Box-Cox transformed random coefficient model (BC-RCM) will be applicable in this context. The model is

$$h(y_{ij},\lambda)=\boldsymbol{x}_{ij}'\boldsymbol{\beta}+\boldsymbol{z}_{ij}'\boldsymbol{b}_i+\varepsilon_{ij},\qquad j=1,\dots,n_i,\quad i=1,\dots,m, \tag{4}$$

where $\boldsymbol{b}_i$ and ε_{ij} are mutually independently distributed as $\boldsymbol{b}_i\sim\mathcal{N}(\mathbf{0},\boldsymbol{R}(\boldsymbol{\phi}))$ and $\varepsilon_{ij}\sim\mathcal{N}(\mathbf{0},\sigma^2)$. The variables and parameters in the form (2) correspond to $\boldsymbol{\psi}=(\boldsymbol{\phi}',\sigma)'$, $\boldsymbol{Z}_i=(\boldsymbol{z}_{i1},\dots,\boldsymbol{z}_{in_i})'$ and $\boldsymbol{G}_i=\sigma^2\boldsymbol{I}_{n_i}$. We apply this mode to the rat growth data in Section 4.3.

2.2 Parameter estimation and its asymptotic properties

Recall that the unknown parameters in the model (2) are $\boldsymbol{\theta}=(\lambda,\boldsymbol{\beta}',\boldsymbol{\psi}')'$. The log-likelihood function of $\boldsymbol{\theta}$ is proportional to

$$\begin{aligned}\ell(\boldsymbol{\theta})=&-\frac{1}{2}\sum_{i=1}^{m}\Big\{\log|\boldsymbol{\Sigma}_i(\boldsymbol{\psi})|+(h(\boldsymbol{y}_i,\lambda)-\boldsymbol{X}_i\boldsymbol{\beta})'\boldsymbol{\Sigma}_i(\boldsymbol{\psi})^{-1}(h(\boldsymbol{y}_i,\lambda)-\boldsymbol{X}_i\boldsymbol{\beta})\Big\}\\&+(\lambda-1)\sum_{i=1}^{m}\sum_{j=1}^{n_i}\log y_{ij}.\end{aligned} \tag{5}$$

When λ and $\boldsymbol{\psi}$ are given, the maximum likelihood (ML) estimator or the generalized least estimator of $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{\psi},\lambda)=\Big\{\sum_{i=1}^{m}\boldsymbol{X}_i'\boldsymbol{\Sigma}_i(\boldsymbol{\psi})^{-1}\boldsymbol{X}_i\Big\}^{-1}\sum_{i=1}^{m}\boldsymbol{X}_i'\boldsymbol{\Sigma}_i(\boldsymbol{\psi})^{-1}h(\boldsymbol{y}_i,\lambda). \tag{6}$$

The parameter $\boldsymbol{\psi}$ is estimated by the ML or the restricted ML estimators. Given λ, we denote an estimator of $\boldsymbol{\psi}$ by $\widehat{\boldsymbol{\psi}}(\lambda)$, which is abbreviated by $\widehat{\boldsymbol{\psi}}$ for notational simplicity. Substituting the estimator $\widehat{\boldsymbol{\psi}}(\lambda)$ into $\widehat{\boldsymbol{\beta}}(\boldsymbol{\psi},\lambda)$, we have $\widehat{\boldsymbol{\beta}}(\lambda)=\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\psi}}(\lambda),\lambda)$.

Concerning the estimation of λ, the ML estimator of λ is derived based on the equation $S(\boldsymbol{\theta}) = 0$, where $S(\boldsymbol{\theta})$ is the partial derivative with respect to λ of the log-likelihood given by

$$S(\boldsymbol{\theta}) = \frac{\partial}{\partial\lambda}\ell(\boldsymbol{\theta}) = -\sum_{i=1}^{m}(h(\boldsymbol{y}_i,\lambda) - \boldsymbol{X}_i\boldsymbol{\beta})'\boldsymbol{\Sigma}_i(\boldsymbol{\psi})^{-1}\left(\frac{\partial}{\partial\lambda}h(\boldsymbol{y}_i,\lambda)\right) + \sum_{i=1}^{m}\sum_{j=1}^{n_i}\log y_{ij},$$

for $(\partial/\partial\lambda)h(y_{ij},\lambda) = (\log y_{ij} - 1/\lambda)h(y_{ij},\lambda) + (1/\lambda)\log y_{ij}$. Since the range of the Box-Cox transformation is not whole real line, the expectation of $S(\boldsymbol{\theta})$ is not zero, which leads to the fact that ML-estimator is inconsistent. Thus, we need an alternative method for estimating λ.

Motivation of applying the Box-Cox transformation to the positive data is to make the transformed data fit normality assumption better than the identity-transformation or the log-transformation. For symmetric distributions including a normal distribution, the third moment is equal to zero. This is why we consider the following moment equation.

$$F(\lambda) \equiv \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left\{h(y_{ij},\lambda) - \boldsymbol{x}'_{ij}\widehat{\boldsymbol{\beta}}(\lambda)\right\}^3 = 0 \tag{7}$$

As a new estimator of λ, we suggest the solution of the equation $F(\lambda) = 0$, which is called here the moment estimator (Mom-estimator) of λ.

When the variance structure of $\boldsymbol{R}$ or $\boldsymbol{G}_i$ is complex, $\widehat{\boldsymbol{\beta}}(\lambda)$ has a complex formula as a function of λ since $\widehat{\boldsymbol{\beta}}(\lambda) = \widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\psi}}(\lambda),\lambda)$ is given in (6). Taking such a case into account, we can implement the following iterative method to get the solution of (7).

Iterative method for getting estimates

1. Determine the initial values $\boldsymbol{\beta}^{(0)}$.
2. Obtain $\lambda^{(1)}$ by solving the equation (7) with $\widehat{\boldsymbol{\beta}}(\lambda) = \boldsymbol{\beta}^{(0)}$.
3. Obtain $\boldsymbol{\psi}^{(1)} = \widehat{\boldsymbol{\psi}}(\lambda^{(1)})$ and $\boldsymbol{\beta}^{(1)} = \widehat{\boldsymbol{\beta}}(\boldsymbol{\psi}^{(1)},\lambda^{(1)})$.
4. Repeat the procedure 2 and 3 until convergence.

A choice of the initial value $\boldsymbol{\beta}^{(0)}$ is the ordinary least squared (OLS) estimator obtained by putting $\boldsymbol{\Sigma}_i = \boldsymbol{I}_{n_i}$ in (6), where we use $\lambda = 0$ (log-transformation) or $\lambda = 1$ (linear transformation) as an initial value of λ. The iterative procedure is easily implemented and performs well as investigated in simulation studies in Section 4. Since $\widehat{\boldsymbol{\beta}}(\lambda) \to \boldsymbol{\beta}$ as $m \to \infty$, it follows from the law of large numbers that

$$m^{-1}F(\lambda) \to E\left[\left\{h(y_{ij},\lambda) - \boldsymbol{x}'_{ij}\boldsymbol{\beta}\right\}^3\right] = 0, \qquad m \to \infty.$$

The fact that the third moment is 0 holds for any symmetric distributions including a normal distribution, whereby we can expect that the Mom-estimator is robust for symmetric distributions far from normality assumption for errors and random effects. We investigate the robustness property of the Mom-estimator by simulation in Section 4.

Based on the Mom-estimator $\widehat{\lambda}$, we can provide the corresponding estimators $\widehat{\boldsymbol{\beta}}(\widehat{\lambda})$ and $\widehat{\boldsymbol{\psi}}(\widehat{\lambda})$ for $\boldsymbol{\beta}$ and $\boldsymbol{\psi}$. To investigate asymptotic properties of these estimators, we assume the following conditions.

Assumption (A)

1. $m^{-1}\sum_{i=1}^{m} \boldsymbol{X}_i'\boldsymbol{\Sigma}_i^{-1}\boldsymbol{X}_i$ converges to a positive definite matrix as $m \to \infty$.

2. The estimator $\widehat{\boldsymbol{\psi}}(\lambda) = (\widehat{\psi}_1(\lambda), \ldots, \widehat{\psi}_k(\lambda))$ is differentiable with respect to λ and satisfies

$$(\widehat{\psi}_j - \psi_j)|\boldsymbol{y}_i = O_p(m^{-1/2}), \quad E(\widehat{\psi}_j - \psi_j|\boldsymbol{y}_i) = O_p(m^{-1}), \quad E((\widehat{\psi}_j - \psi_j)(\widehat{\psi}_l - \psi_l)|\boldsymbol{y}_i) = O_p(m^{-1}),$$
$$d\widehat{\psi}_j(\lambda)/d\lambda = O_p(1), \quad \left\{d\widehat{\psi}_j(\lambda)/d\lambda - E\left[d\widehat{\psi}_j(\lambda)/d\lambda\right]\right\}|\boldsymbol{y}_i = O_p(m^{-1/2})$$

for $j, l = 1, \ldots, k$ and $i = 1, \ldots, m$.

3. For some constant $L > 1$, the eigenvalues of the matrices $\boldsymbol{R}(\boldsymbol{\psi})$ and $\boldsymbol{G}_1(\boldsymbol{\psi}), \ldots, \boldsymbol{G}_m(\boldsymbol{\psi})$ lie in (L^{-1}, L). The eigenvalues of the matrices $\boldsymbol{R}(\widehat{\boldsymbol{\psi}}(\lambda))$ and $\boldsymbol{G}_1(\widehat{\boldsymbol{\psi}}(\lambda)), \ldots, \boldsymbol{G}_m(\widehat{\boldsymbol{\psi}}(\lambda))$ lie in $(L^{-1}/2, 2L)$. The eigenvalues of $\boldsymbol{\Sigma}_1(\boldsymbol{\psi}), \ldots, \boldsymbol{\Sigma}_m(\boldsymbol{\psi})$ lie in a compact set on the positive half of the real line.

4. $\boldsymbol{R}(\boldsymbol{\psi})$ and $\boldsymbol{G}_i(\boldsymbol{\psi})$ has bounded continuous derivatives with respect to $\psi_1, \ldots, \psi_k$.

Then we have the following theorem whose proof is given in the Appendix.

Theorem 1. *Let $\widehat{\lambda}$ be the Mom-estimator given in* (7)*. Under assumption* (A)*, $\widehat{\lambda}$ is consistent, $\widehat{\lambda} - \lambda = O_p(m^{-1/2})$ and $E[\widehat{\lambda} - \lambda] = O(m^{-1})$.*

Theorem 1 shows the consistency of the Mom-estimator $\widehat{\lambda}$ for λ in the Box-Cox transformation, while the ML estimator is not consistent. The consistency of $\widehat{\lambda}$ plays an essential role in constructing a predictor and a prediction interval.

Since $\widehat{\boldsymbol{\beta}}(\lambda)$ and $\widehat{\boldsymbol{\psi}}(\lambda)$ is a smooth function of λ under assumption A, one gets the following corollary immediately.

Corollary 1. *Under assumption* (A)*, $\widehat{\boldsymbol{\beta}}(\widehat{\lambda})$ and $\widehat{\boldsymbol{\psi}}(\widehat{\lambda})$ are consistent, $\widehat{\boldsymbol{\beta}}(\widehat{\lambda}) - \boldsymbol{\beta} = O_p(m^{-1/2})$, $E[\widehat{\boldsymbol{\beta}}(\widehat{\lambda}) - \boldsymbol{\beta}] = O(m^{-1})$, $\widehat{\boldsymbol{\psi}}(\widehat{\lambda}) - \boldsymbol{\psi} = O_p(m^{-1/2})$ and $E[\widehat{\boldsymbol{\psi}}(\widehat{\lambda}) - \boldsymbol{\psi}] = O(m^{-1})$.*

3 Prediction and its Uncertainty

3.1 Transformed predictor

Based on the consistent estimators derived in the previous section, we construct a predictor and a prediction interval for the quantity

$$(\lambda\mu + 1)^{1/\lambda} \quad \text{for} \quad \mu = \boldsymbol{c}_1'\boldsymbol{\beta} + \boldsymbol{c}_2'\boldsymbol{b},$$

where $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$ are known vectors of constants and $\boldsymbol{b} = (\boldsymbol{b}_1', \ldots, \boldsymbol{b}_m')'$. Note that $(\lambda\mu + 1)^{1/\lambda}$ is an inverse function of $h(\mu, \lambda)$ and is interpreted as $\exp(\mu)$ when $\lambda = 0$. The best predictor of $\boldsymbol{b}_i$ in light of the mean squared error is the conditional mean $E[\boldsymbol{b}_i|\boldsymbol{y}_i]$ given by

$$\widehat{\boldsymbol{b}}_i^{\mathrm{BP}}(\boldsymbol{\beta}, \boldsymbol{\psi}, \lambda) = \boldsymbol{R}(\boldsymbol{\psi})\boldsymbol{Z}_i'\boldsymbol{\Sigma}_i(\boldsymbol{\psi})^{-1}\left\{h(\boldsymbol{y}_i, \lambda) - \boldsymbol{X}_i\boldsymbol{\beta}\right\}.$$

When $\boldsymbol{\beta}$, $\boldsymbol{\psi}$ and λ are known, the best predictor of μ is

$$\widehat{\mu}^{\mathrm{BP}}(\boldsymbol{\beta}, \boldsymbol{\psi}, \lambda) = \boldsymbol{c}_1'\boldsymbol{\beta} + \boldsymbol{c}_2'\widehat{\boldsymbol{b}}^{\mathrm{BP}}(\boldsymbol{\beta}, \boldsymbol{\psi}, \lambda), \tag{8}$$

where $\widehat{\boldsymbol{b}}^{\mathrm{BP}} = (\widehat{\boldsymbol{b}}_1^{\mathrm{BP}'}, \ldots, \widehat{\boldsymbol{b}}_m^{\mathrm{BP}'})'$. Thus, the best linear unbiased predictor (BLUP) and the empirical best linear unbiased predictor (EBLUP) are given by

$$\widehat{\mu}^{\mathrm{BLUP}}(\boldsymbol{\psi}, \lambda) = \widehat{\mu}^{\mathrm{BP}}(\widehat{\boldsymbol{\beta}}(\boldsymbol{\psi}, \lambda), \boldsymbol{\psi}, \lambda) = \boldsymbol{c}_1'\widehat{\boldsymbol{\beta}}(\boldsymbol{\psi}, \lambda) + \boldsymbol{c}_2'\widehat{\boldsymbol{b}}^{\mathrm{BLUP}}(\boldsymbol{\psi}, \lambda) \tag{9}$$

for $\widehat{\boldsymbol{b}}^{\mathrm{BLUP}}(\boldsymbol{\psi}, \lambda) = \widehat{\boldsymbol{b}}^{\mathrm{BP}}(\widehat{\boldsymbol{\beta}}(\boldsymbol{\psi}, \lambda), \boldsymbol{\psi}, \lambda)$, and

$$\widehat{\mu}^{\mathrm{EBLUP}} = \widehat{\mu}^{\mathrm{BLUP}}(\widehat{\boldsymbol{\psi}}(\widehat{\lambda}), \widehat{\lambda}) = \boldsymbol{c}_1'\widehat{\boldsymbol{\beta}}(\widehat{\lambda}) + \boldsymbol{c}_2'\widehat{\boldsymbol{b}}^{\mathrm{BLUP}}(\widehat{\boldsymbol{\psi}}(\widehat{\lambda}), \widehat{\lambda}), \tag{10}$$

respectively. From the consistency in Theorem 1 and Corollary 1, it follows that EBLUP is consistent to BLUP as $m \to \infty$.

For predicting the quantity $(\lambda\mu + 1)^{1/\lambda}$, we can inversely transform $\widehat{\mu}^{\mathrm{EBLUP}}$ to get the predictor

$$(\widehat{\lambda}\widehat{\mu}^{\mathrm{EBLUP}} + 1)^{1/\widehat{\lambda}}, \tag{11}$$

which is called the Box-Cox transformed EBLUP (BC-TEBLUP). Since $\widehat{\lambda}$ is consistent, it follows that $(\widehat{\lambda}\widehat{\mu}^{\mathrm{EBLUP}} + 1)^{1/\widehat{\lambda}} \to (\lambda\mu^{\mathrm{BP}} + 1)^{1/\lambda}$ as $m \to \infty$.

3.2 Accurate prediction intervals

We now construct a prediction interval based on BC-TEBLUP given in (11) with a second-order accuracy.

Let $\boldsymbol{Y} = (\boldsymbol{y}_1', \ldots, \boldsymbol{y}_m')'$. The conditional distribution of $\boldsymbol{b}$ given $\boldsymbol{Y}$ is

$$\boldsymbol{b}|\boldsymbol{Y} \sim N(\widehat{\boldsymbol{b}}^{\mathrm{BP}}, \mathrm{diag}\,(\boldsymbol{V}_1, \ldots, \boldsymbol{V}_m)),$$

for $\boldsymbol{V}_i = \boldsymbol{R}(\boldsymbol{\psi}) - \boldsymbol{R}(\boldsymbol{\psi})\boldsymbol{Z}_i'\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\psi})\boldsymbol{Z}_i\boldsymbol{R}(\boldsymbol{\psi})$. Then the conditional distribution of μ given $\boldsymbol{Y}$ is $N(\widehat{\mu}^{\mathrm{BP}}, \sigma_{\mathrm{pos}}^2)$ where σ_{pos} is defined by

$$\sigma_{\mathrm{pos}}^2 = \boldsymbol{c}_2'\mathrm{diag}\,(\boldsymbol{V}_1, \ldots, \boldsymbol{V}_m)\boldsymbol{c}_2, \tag{12}$$

which implies that

$$\sigma_{\mathrm{pos}}^{-1}\left\{h(h^{-1}(\mu, \lambda), \lambda) - \widehat{\mu}^{\mathrm{BP}}\right\}$$

is a standard normal pivot since $h(h^{-1}(\mu, \lambda), \lambda) = \mu$.

To construct a prediction interval of $(\lambda\mu + 1)^{1/\lambda}$, we need to obtain a quantile of the distribution of

$$T = \widehat{\sigma}_{\mathrm{pos}}^{-1}\left\{\frac{(\lambda\mu + 1)^{\widehat{\lambda}/\lambda} - 1}{\widehat{\lambda}} - \widehat{\mu}^{\mathrm{EBLUP}}\right\}, \tag{13}$$

where $\widehat{\sigma}_{\mathrm{pos}}^2$ is the plug-in estimator of σ_{pos}^2. The distribution of T is denoted by $\mathcal{L}_m$. If there were constants a_α and b_α such that $P[a_\alpha \le T_i \le b_\alpha] = 1 - \alpha$, one would get a $100(1 - \alpha)\%$ prediction interval

$$(\lambda\mu + 1)^{1/\lambda} \in \left[\left\{\widehat{\lambda}(\widehat{\mu}^{\mathrm{EBLUP}} + a_\alpha\widehat{\sigma}_{\mathrm{pos}}) + 1\right\}^{1/\widehat{\lambda}},\ \left\{\widehat{\lambda}(\widehat{\mu}^{\mathrm{EBLUP}} + b_\alpha\widehat{\sigma}_{\mathrm{pos}}) + 1\right\}^{1/\widehat{\lambda}}\right].$$

However, $h(h^{-1}(\xi_i,\lambda),\widehat{\lambda})$ is directly affected by randomness of $\widehat{\lambda}$, and the distribution $\mathcal{L}_m$ of (13) depends on unknown parameters. Thus, a_α and b_α are not free from unknown parameters. A feasible approach is an asymptotic approximation of $\mathcal{L}_m$. Since the estimators of the unknown parameters are consistent from Theorem 1 and Corollary 1, it can be seen that $\mathcal{L}_m$ converges to the standard normal distribution under $m\to\infty$. Then, we can get a prediction interval of $(\lambda\mu+1)^{1/\lambda}$ by approximating a_α and b_α with quantiles of the standard normal distribution, However, the accuracy of this prediction interval is of order $O(m^{-1})$, which means that such an approximation does not guarantee enough accuracy when m is moderate.

To achieve second-order accuracy $O(m^{-3/2})$ of prediction intervals, we estimate the distribution $\mathcal{L}_m$ using the parametric bootstrap method. Let $\boldsymbol{Y}^*=((\boldsymbol{y}_1^*)',\ldots,(\boldsymbol{y}_m^*)')'$ be a bootstrap sample which is generated as

$$\boldsymbol{y}_i^*=\left\{1+\widehat{\lambda}(\boldsymbol{X}_i\widehat{\boldsymbol{\beta}}(\widehat{\lambda})+\boldsymbol{Z}_i\boldsymbol{b}_i^*+\boldsymbol{\epsilon}_i^*)\right\}^{1/\widehat{\lambda}},\qquad i=1,\ldots,m$$

where $\boldsymbol{b}_i^*$ and $\boldsymbol{\epsilon}_i^*$ are mutually independently distributed as $\boldsymbol{b}_i^*\sim\mathcal{N}_q(\boldsymbol{0},\boldsymbol{R}(\widehat{\boldsymbol{\psi}}))$ and $\boldsymbol{\epsilon}^*\sim\mathcal{N}_N(\boldsymbol{0},\boldsymbol{G}_i(\widehat{\boldsymbol{\psi}}))$. Based on the bootstrap sample $\boldsymbol{Y}^*$, the estimator $\widehat{\boldsymbol{\beta}}^*(\widehat{\lambda}^*)$, $\widehat{\boldsymbol{\psi}}^*(\widehat{\lambda}^*)$ and $\widehat{\lambda}^*$ are calculated with the same methods as used to obtain $\widehat{\boldsymbol{\beta}}(\widehat{\lambda})$, $\widehat{\boldsymbol{\psi}}(\widehat{\lambda})$ and $\widehat{\lambda}$. For notational simplicity, $\widehat{\boldsymbol{\beta}}^*(\widehat{\lambda}^*)$ and $\widehat{\boldsymbol{\psi}}^*(\widehat{\lambda}^*)$ are abbreviated as $\widehat{\boldsymbol{\beta}}^*$ and $\widehat{\boldsymbol{\psi}}^*$. Let

$$\widehat{\mu}^{\mathrm{EBUP}*}=\widehat{\mu}^{\mathrm{BLUP}}(\widehat{\boldsymbol{\psi}}^*,\widehat{\lambda}^*)=\boldsymbol{c}_1'\widehat{\boldsymbol{\beta}}^*+\boldsymbol{c}_2'\widehat{\boldsymbol{b}}^{\mathrm{BP}}(\widehat{\boldsymbol{\psi}}^*,\widehat{\lambda}^*)$$

and

$$\widehat{\sigma}^{2*}_{\mathrm{pos}}=\boldsymbol{c}_2'\mathrm{diag}\,(\boldsymbol{V}_1(\widehat{\boldsymbol{\psi}}^*),\ldots,\boldsymbol{V}_m(\widehat{\boldsymbol{\psi}}^*))\boldsymbol{c}_2$$

for $\boldsymbol{V}_i(\widehat{\boldsymbol{\psi}}^*)=\boldsymbol{R}(\widehat{\boldsymbol{\psi}}^*)-\boldsymbol{R}(\widehat{\boldsymbol{\psi}}^*)\boldsymbol{Z}_i'\boldsymbol{\Sigma}_i^{-1}(\widehat{\boldsymbol{\psi}}^*)\boldsymbol{Z}_i\boldsymbol{R}(\widehat{\boldsymbol{\psi}}^*)$. Let $\mu^*=\boldsymbol{c}_1'\widehat{\boldsymbol{\beta}}+\boldsymbol{c}_2'\boldsymbol{b}^*$ for $\boldsymbol{b}^*=(\boldsymbol{b}_1^{*'},\ldots,\boldsymbol{b}_m^{*'})'$. Consider the distribution of

$$T^*=(\widehat{\sigma}^*_{\mathrm{pos}})^{-1}\left\{\frac{(\widehat{\lambda}\mu^*+1)^{\widehat{\lambda}^*/\widehat{\lambda}}-1}{\widehat{\lambda}^*}-\widehat{\mu}^{\mathrm{EBLUP}*}\right\},\tag{14}$$

whose distribution is denoted by $\mathcal{L}_m^*$. The approximation accuracy of $\mathcal{L}_m^*$ for $\mathcal{L}_m$ is $O_p(m^{-3/2})$ as shown in the following theorem.

Theorem 2. *Under assumption* (A)*, it follows that*

$$\sup_{q\in\mathbb{R}}|\mathcal{L}_m(q)-\mathcal{L}_m^*(q)|=O_p(m^{-3/2}).\tag{15}$$

The proof of Theorem 2 is given in the Appendix. Using Theorem 2, we obtain a desirable prediction interval of $(\lambda\mu+1)^{1/\lambda}$ with second-order accuracy.

Corollary 2. *For any $\alpha\in(0,1)$, let $q_1=q_1(\boldsymbol{Y})$ and $q_2=q_2(\boldsymbol{Y})$ be quantiles based on the bootstrap sample such that*

$$\mathcal{L}_m^*(q_1)=\alpha/2,\quad \mathcal{L}_m^*(q_2)=1-\alpha/2,$$

where $\mathcal{L}_m^*(\cdot)$ is the distribution function of T^*. Then, the prediction interval of $(\lambda\mu+1)^{1/\lambda}$ with second-order accuracy is given by

$$I_m = \left[\left\{ \widehat{\lambda}(\widehat{\mu}^{\mathrm{EBLUP}} + q_1 \widehat{\sigma}_{\mathrm{pos}}) + 1 \right\}^{1/\widehat{\lambda}}, \ \left\{ \widehat{\lambda}(\widehat{\mu}^{\mathrm{EBLUP}} + q_2 \widehat{\sigma}_{\mathrm{pos}}) + 1 \right\}^{1/\widehat{\lambda}} \right]. \tag{16}$$

In fact, the following holds:

$$P\big((\lambda\mu+1)^{1/\lambda} \in I_m\big) = 1 - \alpha + O(m^{-3/2}). \tag{17}$$

Corollary 2 gives us a highly accurate prediction interval of $(\lambda\xi_i + 1)^{1/\lambda}$ based on BC-TEBLUP. The prediction interval I_m implies that one can figure out precision of BC-TEBLUP with the length of the interval I_m. It is also noted that the coverage accuracy of the prediction interval given in Corollary 2 can be further improved up to $O(m^{-5/2})$ with one round of calibration.

4 Numerical Studies

4.1 Performances of proposed estimators

We begin with investigating finite sample performances of the suggested estimators of the model parameters. In particular, we compare the Mom-estimator proposed in this paper and the ML-estimator which is generally inconsistent.

We generate observations from the model $h(y_{ij}, \lambda) = \beta_0 + \beta_1 x_{ij} + v_i + \varepsilon_{ij}$ for $i = 1, \ldots, m$ with $m = 15$ and $j = 1, \ldots, 5$, where v_i and ε_{ij} are generated from $\mathcal{N}(0, \sigma_1^2)$ and $\mathcal{N}(0, \sigma_2^2)$ with $\sigma_1 = 1$ and $\sigma_2 = 1.5$, respectively, x_{ij} are generated from the uniform distribution on $(4, 8)$, which are fixed through simulation runs, and $\beta_0 = 1$ and $\beta_1 = 2$. The case of $h(y_{ij}, \lambda) = y_{ij}$, that is non-transformed case, corresponds to the nested error regression model proposed in Battese, Harter and Fuller (1988). We consider six patterns of λ's from 0 to 1 by 0.2 and $m = 15$. Based on the $5,000$ Monte Carlo simulation runs, we calculated the root of mean squared errors (RMSE) of estimators of $\lambda, \sigma_v, \sigma_e$ and β_1, defined by

$$\mathrm{RMSE} = \sqrt{E[(\text{estimate} - \text{true value})^2]}.$$

For estimation of λ, we use the Mom-estimator given in (7) and the ML-estimator. For estimation of σ_v and σ_e, we use the maximum likelihood estimator. The resulting RMSE for each value of λ is given in Figure 1. In all cases of λ, the RMSEs in the ML-estimator (dotted line) are considerably larger than those in the Mom-estimator, and the Mom-estimator provides stable estimates regardless of λ. It is also observed that the RMSEs in the ML-estimator gets large as λ increases. Thus, Figure 1 shows that the Mom-estimator yields remarkable improvement over the ML-estimator in terms of EMSE.

We next investigate robustness properties in estimation of λ. Following the setup in Datta, Rao and Smith (2005), we replace both random effects $\mathcal{N}(0, \sigma_1^2)$ and error components $\mathcal{N}(0, \sigma_2^2)$ by the three distribution: t-distribution with 3 degrees of freedom, Laplace (double exponential) distribution and location exponential distribution. The variance of three distributions are set

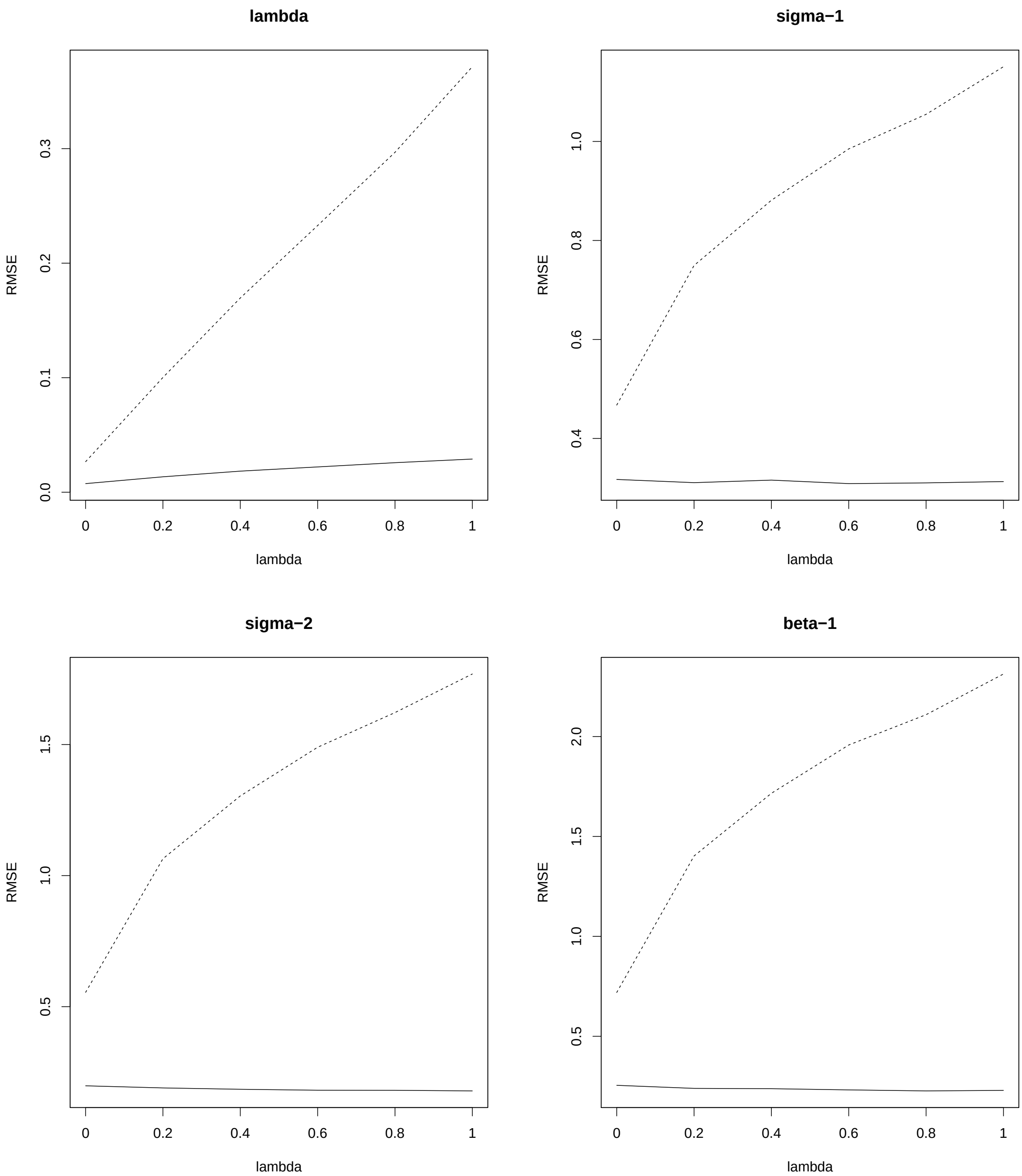

Figure 1: Simulated Root Mean Squared Errors (RMSE) of Estimators of Model Parameters (Solid line corresponds to the proposed Mom-estimator and dotted line the conventional ML-estimator).

to be the same as σ_1^2 and σ_2^2. We consider the same data generating process with $\lambda = 0.5$ and $m = 15, 30$ other than the distribution of v_i's and ε_{ij}'s. Based on the 5,000 simulation runs, we calculate the mean values and standard deviations of the estimators of the transformation parameter λ. The results are reported in Table 1. It is observed that the averages of estimates of both ML- and Mom-estimators in t-distribution and Laplace distribution are close to the true value $\lambda = 0.5$, but the standard deviations of the ML-estimator are much larger than those of the Mom-estimator. In the case of location exponential distribution, the average of estimates of the ML-estimator is far away from the true value, whereas the Mom-estimator provides rather good estimates with small standard deviations. Also the standard deviations in the normal case are always smaller than the other cases. Thus, it is concluded that the Mom-estimator has a nice robustness property compared to the conventional ML-estimator.

Table 1: Average Values and Standard Errors of Estimators of λ Under Misspecified Distribution (True value is $\lambda = 0.5$).

		ML-estimator		Mom-estimator	
	true distribution	mean	sd	mean	sd
m=15	Normal	0.48	0.19	0.50	0.02
	Student-t	0.47	0.27	0.50	0.08
	Laplace	0.49	0.24	0.50	0.04
	Location exponential	0.05	0.23	0.43	0.03
m=30	Normal	0.49	0.14	0.50	0.01
	Student-t	0.48	0.21	0.50	0.07
	Laplace	0.50	0.18	0.50	0.03
	Location exponential	0.06	0.19	0.43	0.02

4.2 Performances of prediction intervals

We here investigate finite sample performances of the suggested prediction interval based on parametric bootstrap. We use the same data generating process given in previous subsection with $\lambda = 0.5, 1$ and $m = 10$. The frequency of the prediction intervals which includes $h^{-1}(\xi_i, \lambda)$ is counted for $i = 1, \ldots, m$, and the coverage probability is estimated by dividing the total number of the frequency by 1,000, where the size of the bootstrap sample is 200. The expected length of the prediction interval can be also estimated as an average length by a similar method.

Under the above simulation settings, we compare the suggested prediction interval with the naive prediction interval given by

$$\left[\left\{ \widehat{\lambda}(\widehat{\mu}^{\mathrm{EBLUP}} - z_{\alpha/2}\widehat{\sigma}_{\mathrm{pos}}) + 1 \right\}^{1/\widehat{\lambda}}, \ \left\{ \widehat{\lambda}(\widehat{\mu}^{\mathrm{EBLUP}} + z_{\alpha/2}\widehat{\sigma}_{\mathrm{pos}}) + 1 \right\}^{1/\widehat{\lambda}} \right], \tag{18}$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile point of the standard normal distribution. This is an empirical Bayes confidence interval which is derived by substituting the estimators into the Bayes confidence interval. Table 2 reports the coverage probability (CP) and the expected length (EL) of the two prediction intervals (16) and (18) based on the bootstrap method (BT) and the naive method (NV) with nominal confidence coefficients $1 - \alpha = 0.95$ and $1 - \alpha = 0.99$. From Table 2, we can observe that the naive prediction interval is not appropriate since it does not satisfy the nominal confidence coefficient, while it gives a shorter length than BT. On the other hand, the prediction interval (16) based on BT has the coverage probability close to the nominal level 0.95 and 0.99, respectively, which shows that the correction by the bootstrap method works well.

Table 2: Values of Coverage Probability and Expected Length of the Unconditional Prediction Interval with Confidence Coefficient $1-\alpha=0.95$ and $1-\alpha=0.99$.

λ		0.5		1	
		NV	BT	NV	BT
$\alpha = 0.05$	CP	89.1	97.0	90.0	97.5
	EL	14.8	102.7	2.0	11.6
$\alpha = 0.01$	CP	94.8	99.0	95.3	99.2
	EL	19.8	343.2	2.65	26.0

4.3 Application to rat growth data

We now apply the proposed model to the rat growth data treated in Gelfand et al.(1990), where 60 different rats are weighted at five different points in time, and 60 rats are equally divided into control and treatment groups. The observations of weights are denoted by $y_{ij}^{(k)}, i = 1,\ldots,30,\ j = 1,\ldots,5$, where $k = 1,2$ denotes the control group and treatment group, respectively. Let $t_j,\ j = 1,\ldots,5$ be the measurement time of weighting. We apply the Box-Cox transformed random coefficient model described as

$$h(y_{ij}^{(k)}, \lambda^{(k)}) = \beta_0^{(k)} + \beta_1^{(k)} t_j + u_i^{(k)} + v_i^{(k)} t_j + \varepsilon_{ij}, \tag{19}$$

where $u_i^{(k)}$ and $v_i^{(k)}$ are mutually independent and distributed $u_i^{(k)} \sim N(0, \tau^{2(k)})$, $v_i^{(k)} \sim N(0, \sigma^{2(k)})$. We used the maximum likelihood estimator for variance components $\tau^{2(k)}$ and $\sigma^{2(k)}$. The estimates of parameters are given by

	$\widehat{\lambda}$	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\hat{\tau}^2$	$\widehat{\sigma}^2$
Control ($k=1$)	0.248	9.32	0.103	0.008	0.209
Treatment ($k=2$)	0.270	9.54	0.103	0.014	0.238

It is observed from the estimated values of λ that nearly biquadratic root transformation is appropriate for applying the rat growth data to a linear mixed model. The estimated growth functions in both treatment and control groups, namely

$$f^{(k)}(t_j) = \{1 + \widehat{\lambda}^{(k)}(\widehat{\beta}_0^{(k)} + \widehat{\beta}_1^{(k)} t_j)\}^{1/\widehat{\lambda}^{(k)}}, \quad k = 1,2, \quad j = 1,\ldots,5$$

are given in Figure 2. It is interesting to point out that the difference of growth curves between the control and treatment groups appears in the estimates of $\widehat{\lambda}$. Actually, it is observed that the

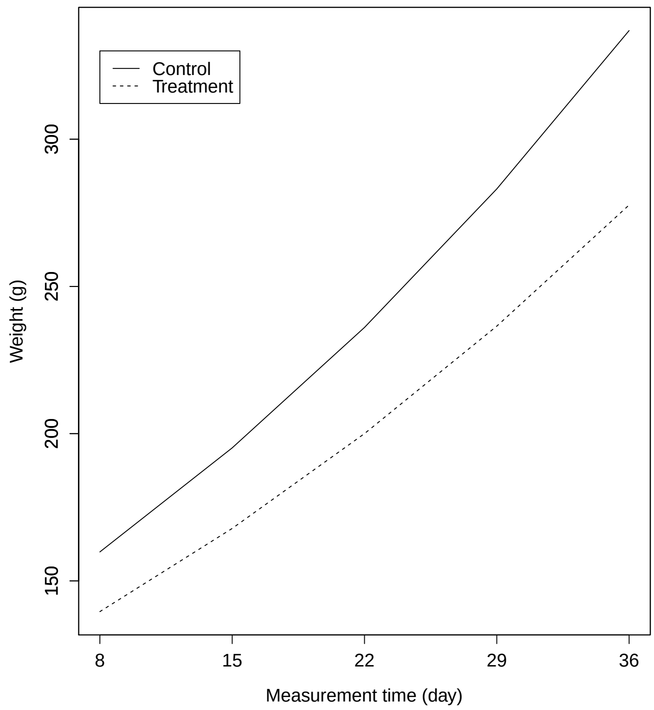

Figure 2: Estimated Growth Function by BC-LMM for Rat Growth Data

growth curve of the treatment group is more rapidly growing than the control group. This is because the estimates of the transformation parameter λ in the treatment group is $\widehat{\lambda}^{(2)} = 0.270$, which bigger than the estimates $\widehat{\lambda}^{(1)} = 0.248$ in the control group.

We next construct prediction intervals using the data. We here treat the problem of predicting a future growth pattern for each rat, where the weight of the first time t_1 only is observed. That is, we construct the prediction intervals of the following quantity

$$\left\{1 + \lambda^{(k)}(\beta_0^{(k)} + \beta_1^{(k)} t_j + u_i^{(k)} + v_i^{(k)})\right\}^{1/\lambda^{(k)}}, \quad k = 1, 2, \quad j = 2, \ldots, 5.$$

The prediction intervals derived for 2-nd rat in the control group and 16-th rat in the treatment group are given in Figure 3. From the Figure, the proposed parametric bootstrap method works well in this situation. It is also observed that the length of prediction intervals gets large as t_j increases. This is consistent to the fact that the variability of the observed growth curves extends as t_j increases.

4.4 Application to PLP data in Japan

We next apply the Box-Cox transformed nested error regression model (BC-TNERM) defined in (3) to the posted land price (PLP) data along the Keikyu train line, which connects the suburbs in Kanagawa prefecture to the Tokyo metropolitan area. This data set was used by Kawakubo and Kubokawa (2014). We analyze the land price data in 2001 with covariates for

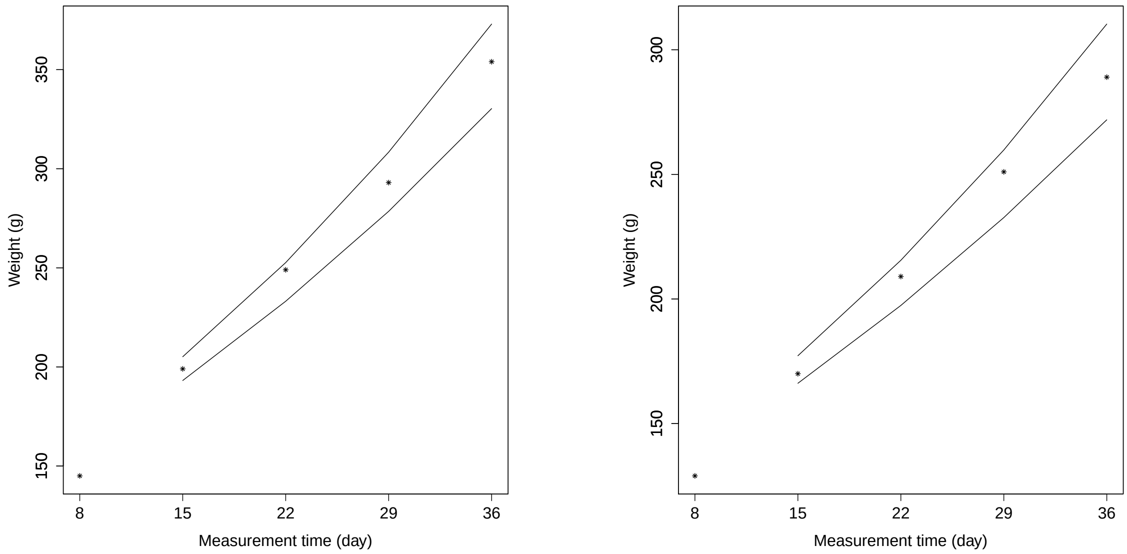

Figure 3: Prediction Intervals for Future Growth Patterns (right: 2-nd rat in a control group, left: 16-th rat in a treatment group and the asterisk means the actual observation).

51 stations which we consider as small areas with $m = 51$. For the i-th small area, there are data of n_i land spots, and the total sample size is $N = \sum_{j=1}^{m} n_i = 192$. The land price (Yen in hundreds of thousands) per m^2 of the j-th spot in the i-th small area is denoted by y_{ij} (scaled by 1000), TRN_i is the time to take by train from the station i to the Tokyo station around 9:00 in the morning, DST_{ij} is the geographical distance from the spot j to the nearby station i and FAR_{ij} denotes the floor-area ratio of the spot j. As explanatory variables, we consider $\log(\mathrm{FAR}_{ij})$, $\log(\mathrm{TRN}_i)$ and $\log(\mathrm{DST}_{ij})$. Then we consider the following model:

$$h(y_{ij}, \lambda) = \beta_0 + \beta_1 \log(\mathrm{FAR}_{ij}) + \beta_2 \log(\mathrm{TRN}_i) + \beta_3 \log(\mathrm{DST}_{ij}) + v_i + \varepsilon_{ij}, \tag{20}$$

where v_i and ε_{ij} are mutually independently distributed as $v_i \sim \mathcal{N}(0, \tau^2)$ and $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$. We used the maximum likelihood estimator for variance components τ^2 and σ^2. The estimated values are

$$\widehat{\lambda} = 0.43,\ \hat{\tau} = 0.67,\ \widehat{\sigma} = 2.21,\ \widehat{\beta}_0 = 49.0,\ \widehat{\beta}_1 = 2.05,\ \widehat{\beta}_2 = -7.10,\ \widehat{\beta}_3 = -1.18.$$

It comes from the result that the suitable transformation for this data set is either log-transformation ($\lambda = 0$) nor linear transformation ($\lambda = 1$), but the approximately squared transformation ($\lambda = 0.5$) is an appropriate choice. Based on the 1000 bootstrap samples, we construct the prediction intervals given in Figure 4. We see that our bootstrap method seems to work well and provides reasonable prediction intervals.

5 Concluding remarks

In this paper, we treat the linear mixed models with the Box-Cox transformation. Based on the moment method, we have suggested the simple and consistent estimator of the transformation parameter λ. The consistent estimator has provided the reasonable predictor and the

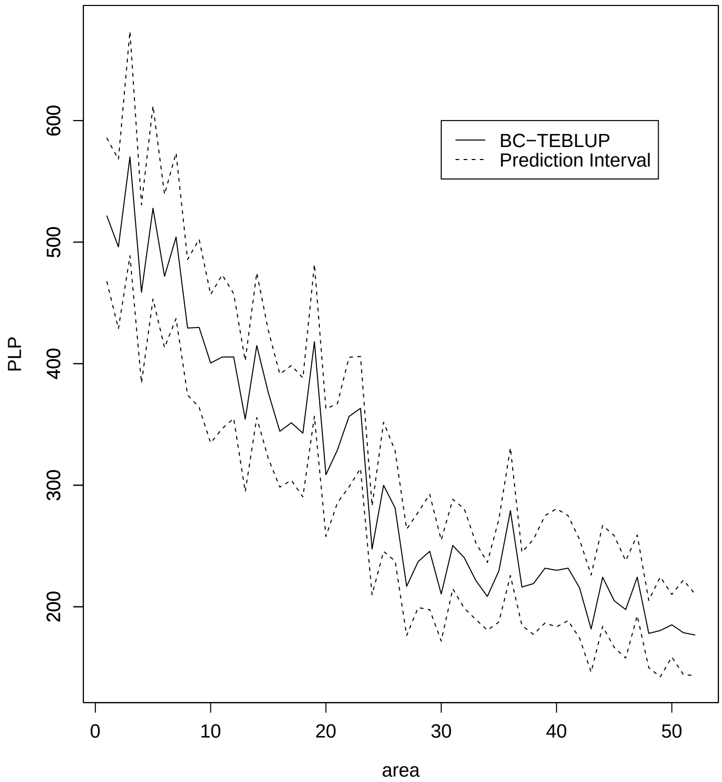

Figure 4: Values of TBLUP and its Prediction Interval for PLP Data.

prediction interval with second-order accuracy. Through the simulation and empirical study, we have confirmed that the estimators of the parameters and the prediction interval have good performances and the Box-Cox linear mixed model is applicable.

Although we have considered the prediction problem in this paper, the Box-Cox transformation can be used in various problems including the variable selection in linear regression models, the testing hypothesis for linear constraint on regression coefficients and others. The consistent estimation methods proposed in this paper will be applicable to these problems.

Appendix

All the technical parts including the proofs of lemmas and theorems are given here.

A.1 Proof of Theorem 1. We begin by demonstrating the consistency of $\widehat{\lambda}$. According the Cramer method explained in Jiang (2010), we show that the equation $F(\lambda)=0$ includes a solution which converges to λ in probability. Let

$$g_m(\lambda')=m^{-1}F(\lambda'),$$

for scalar λ'. Then, it can be seen that $g_m(\lambda')$ converges to $g(\lambda')$ in probability, where

$$g(\lambda') = \lim_{m\to\infty} m^{-1}E_\lambda[F(\lambda')].$$

When $\lambda' = \lambda$, it is noted that $g(\lambda) = 0$, since $g(\lambda) = \lim_{m\to\infty} m^{-1}E_\lambda[F(\lambda)] = 0$. Since $g(\lambda')$ is continuous, without loss of generality, we have $g(\lambda - \varepsilon) < 0$ and $g(\lambda + \varepsilon)$ for some positive ε. Then, $g_m(\lambda-\varepsilon)$ and $g_m(\lambda+\varepsilon)$ converge to $g(\lambda-\varepsilon) < 0$ and $g(\lambda+\varepsilon)$, respectively, in probability. This implies that both probabilities $P(g_m(\lambda - \varepsilon) < 0)$ and $P(g_m(\lambda + \varepsilon) > 0)$ converge to one as $N \to \infty$. In fact, for instance, the former result follows from the fact that

$$\begin{aligned}P(g_m(\lambda-\varepsilon) < 0) =& P(g_m(\lambda-\varepsilon) - g(\lambda-\varepsilon) < -g(\lambda-\varepsilon)) \\ >& P(|g_m(\lambda-\varepsilon) - g(\lambda-\varepsilon)| < -g(\lambda-\varepsilon)) \to 1,\end{aligned}$$

as $m \to \infty$ since $-g(\lambda - \varepsilon) > 0$. Thus, for any $\delta > 0$, there exists an M such that for any $m > M$, $P(g_m(\lambda - \varepsilon) < 0) > 1 - \delta$ and $P(g_m(\lambda + \varepsilon) > 0) > 1 - \delta$. Note that the intersection of the events $\{g_m(\lambda - \varepsilon) < 0\}$ and $\{g_m(\lambda + \varepsilon) > 0\}$ implies that $\widehat{\lambda}$ is included in the interval $(\lambda - \varepsilon, \lambda + \varepsilon)$, namely, $|\widehat{\lambda} - \lambda| < \varepsilon$. Hence,

$$P(|\widehat{\lambda} - \lambda| < \varepsilon) > P(g_m(\lambda - \varepsilon) < 0,\ g_m(\lambda + \varepsilon) < 0) > 1 - 2\delta,$$

which means that $\widehat{\lambda}$ is consistent.

We next show that $\widehat{\lambda} - \lambda = O_p(m^{-1/2})$. To this end, we expand the equation (7) around λ to get

$$\sqrt{m}(\widehat{\lambda} - \lambda) = -\frac{m^{-1/2}F(\lambda)}{m^{-1}F_\lambda(\lambda^*)}, \tag{21}$$

where λ^* is an intermediate value between λ and $\widehat{\lambda}$, and $F_\lambda(\lambda) = \partial F(\lambda)/\partial\lambda$. It follows that

$$F_\lambda(\lambda^*) = 3\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left\{h(y_{ij},\lambda) - \boldsymbol{x}_{ij}'\widehat{\boldsymbol{\beta}}(\lambda)\right\}^2\left\{h_\lambda(y_{ij},\lambda) - \boldsymbol{x}_{ij}'\widehat{\boldsymbol{\beta}}_\lambda(\lambda)\right\},$$

where $\widehat{\boldsymbol{\beta}}_\lambda(\lambda) = \partial\widehat{\boldsymbol{\beta}}(\lambda)/\partial\lambda$. Since

$$h_\lambda(y_{ij},\lambda) = \frac{y_{ij}^\lambda \log y_{ij}^\lambda - y_{ij}^\lambda + 1}{\lambda^2} = \frac{y_{ij}^\lambda \log y_{ij}^\lambda - \lambda h(y_{ij},\lambda)}{\lambda^2}$$

and $-e^{-1} \le y_{ij}^\lambda \log y_{ij}^\lambda \le y_{ij}^{2\lambda} = (1 + \lambda h(y_{ij},\lambda))^2$, we have

$$\left|\left\{h(y_{ij},\lambda) - \boldsymbol{x}_{ij}'\boldsymbol{\beta}\right\}^2 h_\lambda(y_{ij},\lambda)\right| \le \frac{1}{\lambda^2}\left\{h(y_{ij},\lambda) - \boldsymbol{x}_{ij}'\boldsymbol{\beta}\right\}^2\left(\max\left\{e^{-1},(1+\lambda h(y_{ij},\lambda))^2\right\} + |\lambda h(y_{ij},\lambda)|\right),$$

whereby

$$E\left|\left\{h(y_{ij},\lambda) - \boldsymbol{x}_{ij}'\boldsymbol{\beta}\right\}^2 h_\lambda(y_{ij},\lambda)\right| < \infty.$$

Moreover, we have

$$\begin{aligned}&m^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left\{h(y_{ij},\lambda) - \boldsymbol{x}_{ij}'\boldsymbol{\beta}\right\}^2\boldsymbol{x}_{ij}'\widehat{\boldsymbol{\beta}}_\lambda(\lambda) \\ &= \left(m^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left\{h(y_{ij},\lambda) - \boldsymbol{x}_{ij}'\boldsymbol{\beta}\right\}^2\boldsymbol{x}_{ij}'\right)\widehat{\boldsymbol{\beta}}_\lambda(\lambda) = O_p(1)\end{aligned}$$

since $E\left|\widehat{\beta}_{\lambda,i}(\lambda)\right| < \infty$, where $\widehat{\beta}_{\lambda,i}(\lambda)$ denotes the i-th element in $\widehat{\boldsymbol{\beta}}_{\lambda}(\lambda)$, which will be shown in the end of the proof. It follows from $\widehat{\boldsymbol{\beta}}(\lambda) = \boldsymbol{\beta} + O_p(m^{-1/2})$ and consistency of $\widehat{\lambda}$ that

$$m^{-1}F_{\lambda}(\lambda^*) = m^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left\{h(y_{ij},\lambda) - \boldsymbol{x}_{ij}'\boldsymbol{\beta}\right\}^2\left\{h_{\lambda}(y_{ij},\lambda) - \boldsymbol{x}_{ij}'\widehat{\boldsymbol{\beta}}_{\lambda}(\lambda)\right\} + o_p(1),$$

which indicates that

$$m^{-1}F_{\lambda}(\lambda^*) = O_p(1)$$

since

$$\sum_{j=1}^{n_i}\left\{h(y_{ij},\lambda) - \boldsymbol{x}_{ij}'\boldsymbol{\beta}(\lambda)\right\}^2 h_{\lambda}(y_{ij},\lambda)$$

is independent for each i. For denominator in (21), it is observed that

$$m^{-1/2}F(\lambda) = m^{-1/2}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left\{h(y_{ij},\lambda) - \boldsymbol{x}_{ij}'\boldsymbol{\beta}\right\}^3 + o_p(1),$$

$E\left\{h(y_{ij},\lambda) - \boldsymbol{x}_{ij}'\boldsymbol{\beta}\right\}^3 = 0$ and $E\left\{h(y_{ij},\lambda) - \boldsymbol{x}_{ij}'\boldsymbol{\beta}\right\}^6 < \infty$ since $h(y_{ij},\lambda)$ is normally distributed. Then from the independence of $\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left\{h(y_{ij},\lambda) - \boldsymbol{x}_{ij}'\boldsymbol{\beta}\right\}^3$, we have from the central limit theorems that

$$m^{-1/2}F(\lambda) = O_p(1),$$

whereby $\sqrt{m}(\widehat{\lambda} - \lambda) = O_p(1)$, namely $\widehat{\lambda} - \lambda = O_p(m^{-1/2})$.

For the proof of $E[\widehat{\lambda} - \lambda] = O(m^{-1})$, we begin with approximating $\widehat{\lambda} - \lambda$ stochastically

$$\widehat{\lambda} - \lambda = -\frac{F(\lambda)}{F_{\lambda}(\lambda)} + O_p(m^{-1}) = -\frac{F(\lambda)}{E\left\{F_{\lambda}(\lambda)\right\}} + O_p(m^{-1}),$$

where $E\left[F(\lambda)\right] = O(m)$. Since

$$\begin{aligned} E\{F(\lambda)\} &= \sum_{i=1}^{m}\sum_{j=1}^{n_i} E\left\{h(y_{ij},\lambda) - \boldsymbol{x}_{ij}'\boldsymbol{\beta} - \boldsymbol{x}_{ij}'(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right\}^3 \\ &= \sum_{i=1}^{m}\sum_{j=1}^{n_i} E\left\{h(y_{ij},\lambda) - \boldsymbol{x}_{ij}'\boldsymbol{\beta}\right\}^3 + O(1) = O(1), \end{aligned}$$

since $E\left\{h(y_{ij},\lambda) - \boldsymbol{x}_{ij}'\boldsymbol{\beta}\right\}^3 = 0$, $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = O_p(m^{-1/2})$ and $E(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = O(m^{-1})$. Thus, we obtain $E[\widehat{\lambda} - \lambda] = O(m^{-1})$.

Finally, we establish that $E\left|\widehat{\beta}_{\lambda,i}(\lambda)\right| < \infty$ for $i = 1, \ldots, p$. From the expression of $\widehat{\boldsymbol{\beta}}(\boldsymbol{\psi},\lambda)$ given in (6), we have

$$\begin{aligned} \frac{\partial\widehat{\boldsymbol{\beta}}(\lambda)}{\partial\lambda} &= \left(\sum_{i=1}^{m}\boldsymbol{X}_i'\boldsymbol{\Sigma}_i(\widehat{\boldsymbol{\psi}})^{-1}\boldsymbol{X}_i\right)^{-1}\left(\sum_{i=1}^{m}\boldsymbol{X}_i'\boldsymbol{\Sigma}_i(\widehat{\boldsymbol{\psi}})^{-1}\boldsymbol{\Sigma}_{i(\lambda)}(\widehat{\boldsymbol{\psi}})\boldsymbol{\Sigma}_i(\widehat{\boldsymbol{\psi}})^{-1}\boldsymbol{X}_i\right)(\widehat{\boldsymbol{\beta}}(\lambda) - \widehat{\boldsymbol{\beta}}^*(\lambda)) \\ &\quad + \left(\sum_{i=1}^{m}\boldsymbol{X}_i'\boldsymbol{\Sigma}_i(\widehat{\boldsymbol{\psi}})^{-1}\boldsymbol{X}_i\right)^{-1}\sum_{i=1}^{m}\boldsymbol{X}_i'\boldsymbol{\Sigma}_i(\widehat{\boldsymbol{\psi}})^{-1}h_{\lambda}(\boldsymbol{y}_i,\lambda), \end{aligned}$$

where

$$\widehat{\boldsymbol{\beta}}^*(\lambda) = \left(\sum_{i=1}^m \boldsymbol{X}_i' \boldsymbol{\Sigma}_i(\widehat{\boldsymbol{\psi}})^{-1} \boldsymbol{\Sigma}_{i(\lambda)}(\widehat{\boldsymbol{\psi}}) \boldsymbol{\Sigma}_i(\widehat{\boldsymbol{\psi}})^{-1} \boldsymbol{X}_i \right)^{-1} \sum_{i=1}^m \boldsymbol{X}_i' \boldsymbol{\Sigma}_i(\widehat{\boldsymbol{\psi}})^{-1} \boldsymbol{\Sigma}_{i(\lambda)}(\widehat{\boldsymbol{\psi}}) \boldsymbol{\Sigma}_i(\widehat{\boldsymbol{\psi}})^{-1} h(\boldsymbol{y}_i, \lambda)$$

with

$$\boldsymbol{\Sigma}_{i(\lambda)}(\widehat{\boldsymbol{\psi}}) = \frac{\partial}{\partial \lambda} \boldsymbol{\Sigma}_i(\widehat{\boldsymbol{\psi}}) = \sum_{l=1}^k \left(\frac{\partial}{\partial \psi_l} \boldsymbol{\Sigma}_i(\boldsymbol{\psi}) \right) \frac{\partial \widehat{\boldsymbol{\psi}}_l(\lambda)}{\partial \lambda},$$

which is $O_p(1)$ under condition 2 in assumption A. Since $\widehat{\psi}_l - \psi_l = O_p(m^{-1/2}),\ l = 1, \ldots, k$ and $\boldsymbol{\Sigma}_i(\boldsymbol{\psi})$ is a smooth function of $\boldsymbol{\psi}$, it follows that

$$\begin{aligned} &\frac{1}{m} \sum_{i=1}^m \boldsymbol{X}_i' \boldsymbol{\Sigma}_i(\widehat{\boldsymbol{\psi}})^{-1} \boldsymbol{\Sigma}_{i(\lambda)}(\widehat{\boldsymbol{\psi}}) \boldsymbol{\Sigma}_i(\widehat{\boldsymbol{\psi}})^{-1} h(\boldsymbol{y}_i, \lambda) \\ &\quad = \frac{1}{m} \sum_{i=1}^m \sum_{l=1}^k \boldsymbol{X}_i' \boldsymbol{\Sigma}_i(\boldsymbol{\psi})^{-1} \left(\frac{\partial}{\partial \psi_l} \boldsymbol{\Sigma}_i(\boldsymbol{\psi}) \right) \boldsymbol{\Sigma}_i(\boldsymbol{\psi})^{-1} \left(h(\boldsymbol{y}_i, \lambda) \frac{\partial \widehat{\boldsymbol{\psi}}_l(\lambda)}{\partial \lambda} \right) + O_p(m^{-1/2}). \end{aligned}$$

It is noted that

$$\begin{aligned} h(\boldsymbol{y}_i, \lambda) \cdot \frac{\partial \widehat{\boldsymbol{\psi}}_l(\lambda)}{\partial \lambda} &= \left(\frac{\partial \widehat{\boldsymbol{\psi}}_l(\lambda)}{\partial \lambda} - E\left[\frac{\partial \widehat{\boldsymbol{\psi}}_l(\lambda)}{\partial \lambda} \right] \right) h(\boldsymbol{y}_i, \lambda) + E\left[\frac{\partial \widehat{\boldsymbol{\psi}}_l(\lambda)}{\partial \lambda} \right] h(\boldsymbol{y}_i, \lambda) \\ &= E\left[\frac{\partial \widehat{\boldsymbol{\psi}}_l(\lambda)}{\partial \lambda} \right] h(\boldsymbol{y}_i, \lambda) + O_p(m^{-1/2}) \end{aligned}$$

under condition 2 in assumption A. Thus we get

$$\begin{aligned} \sqrt{m} \left(\widehat{\boldsymbol{\beta}}^*(\lambda) - \boldsymbol{\beta} \right) &= \left(\frac{1}{m} \sum_{i=1}^m \boldsymbol{X}_i' \boldsymbol{\Sigma}_i(\boldsymbol{\psi})^{-1} E\left[\boldsymbol{\Sigma}_{i(\lambda)}(\widehat{\boldsymbol{\psi}}) \right] \boldsymbol{\Sigma}_i(\boldsymbol{\psi})^{-1} \boldsymbol{X}_i \right)^{-1} \\ &\quad \times \frac{1}{\sqrt{m}} \sum_{i=1}^m \boldsymbol{X}_i' \boldsymbol{\Sigma}_i(\boldsymbol{\psi})^{-1} E\left[\boldsymbol{\Sigma}_{i(\lambda)}(\widehat{\boldsymbol{\psi}}) \right] \boldsymbol{\Sigma}_i(\boldsymbol{\psi})^{-1} \left(h(\boldsymbol{y}_i, \lambda) - \boldsymbol{X}_i \boldsymbol{\beta} \right) + O_p(1), \end{aligned}$$

which is $O_p(1)$ by the central limit theorem and independence of $\boldsymbol{y}_i,\ i = 1, \ldots, m$. Therefore, $\widehat{\boldsymbol{\beta}}^*(\lambda) - \boldsymbol{\beta} = O_p(m^{-1/2})$ follows, whereby $\widehat{\boldsymbol{\beta}}(\lambda) - \widehat{\boldsymbol{\beta}}^*(\lambda) = O_p(m^{-1/2})$. Using the similar argument and the law of large numbers, we have $E\left| \widehat{\beta}_{\lambda,i}(\lambda) \right| = O(1)$. □

A.2 Proof of Theorem 2. Let $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}(\widehat{\lambda})', \boldsymbol{\psi}(\widehat{\lambda}), \widehat{\lambda})'$ the estimator of $\boldsymbol{\theta}$ given in Section 2. If $\mathcal{L}_m(q)$ is expanded as

$$\mathcal{L}_m(q) = \Phi(q) + m^{-1} \gamma(q, \boldsymbol{\theta}) + O_p(m^{-3/2}), \tag{22}$$

where $\gamma(q, \boldsymbol{\theta})$ is a smooth function with $O(1)$, then the corresponding expansion holds for $\mathcal{L}_m^*(q)$, namely,

$$\mathcal{L}_m^*(q) = \Phi(q) + m^{-1} \gamma(q, \widehat{\boldsymbol{\theta}}) + O_p(m^{-3/2}).$$

Thus, one gets

$$
\begin{aligned}
\mathcal{L}_m^*(q)-\mathcal{L}_m(q)=&m^{-1}\big\{\gamma(q,\widehat{\boldsymbol{\theta}})-\gamma(q,\boldsymbol{\theta})\big\}+O_p(m^{-3/2})\\
=&m^{-1}\Big(\frac{\partial\gamma(q,\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\Big)'(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta})+O_p(m^{-3/2}),
\end{aligned}
$$

which establishes the result given in Theorem 2. Since the inequality

$$
\widehat{\sigma}_{\text{pos}}^{-1}\left\{\frac{(\lambda\mu+1)^{\widehat{\lambda}/\lambda}-1}{\widehat{\lambda}}-\widehat{\mu}^{\text{EBLUP}}\right\}\leq q
$$

for any $q\in\mathbb{R}$ is equivalently rewritten as

$$
(\lambda\mu+1)^{1/\lambda}\leq\left\{\widehat{\lambda}(\widehat{\mu}^{\text{EBLUP}}+q\widehat{\sigma}_{\text{pos}})+1\right\}^{1/\widehat{\lambda}},
$$

we have

$$
\begin{aligned}
\mathcal{L}_m(q)&=P\left[\widehat{\sigma}_{\text{pos}}^{-1}\left\{\frac{(\lambda\mu+1)^{\widehat{\lambda}/\lambda}-1}{\widehat{\lambda}}-\widehat{\mu}^{\text{EBLUP}}\right\}\leq q\right]\\
&=E\left(P\left[\sigma_{\text{pos}}^{-1}(\mu-\widehat{\mu}^{\text{BP}})\leq\sigma_{\text{pos}}^{-1}\left\{\frac{(\widehat{\lambda}(\widehat{\mu}^{\text{EBLUP}}+q\widehat{\sigma}_{\text{pos}})+1)^{\lambda/\widehat{\lambda}}-1}{\lambda}\right\}\right]\Big|\boldsymbol{Y}\right)\\
&=E\left[\Phi(q+R(q,\boldsymbol{Y}))\right]
\end{aligned}
$$

where $\Phi(\cdot)$ is a cumulative distribution function of the standard normal distribution and

$$
R(q,\boldsymbol{Y})=\frac{1}{\sigma_{\text{pos}}\lambda}\Big[\Big\{\widehat{\lambda}(\widehat{\mu}^{\text{EBLUP}}+q\sigma_{\text{pos}})+1\Big\}^{\lambda/\widehat{\lambda}}-1\Big]-q.
$$

For the standard normal density function $\phi(\cdot)$, the Taylor expansion is applied to get

$$
\mathcal{L}_m(q)=\Phi(q)+\phi(q)t_1(q)-\frac{1}{2}q\phi(q)t_2(q)+t_3(q),
$$

where $t_1(q)=E[R(q,\boldsymbol{Y})],t_2(q)=E[R^2(q,\boldsymbol{Y})]$ and

$$
t_3(q)=\frac{1}{2}E\left[\int_q^{q+R(q,\boldsymbol{Y})}\{q+R(q,\boldsymbol{Y})-x\}^2(x^2-1)dx\right]\leq C_1E|R(q,\boldsymbol{Y})|^3
$$

for some constant C_1. Since $\boldsymbol{y}_1,\ldots,\boldsymbol{y}_m$ are mutually independent and it is easy to get the result that $\widehat{\lambda}-\lambda|\boldsymbol{y}_i=O_p(m^{-1/2})$ and $E(\widehat{\lambda}-\lambda|\boldsymbol{y}_i)=O_p(m^{-1})$ for $i=1,\ldots,m$, which deduce $(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta})|\boldsymbol{y}_i=O_p(m^{-1/2})$ and $E(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}|\boldsymbol{y}_i)=O_p(m^{-1})$ for $i=1,\ldots,m$. Then we obtain that $t_1(q)$ and $t_2(q)$ is of order $O(m^{-1})$ by the similar way given in the proof of Theorem 1 in Sugasawa and Kubokawa (2014b), whereby the formula (22) is established. □

References

[1] Baltagi, B. H. (2013). *Econometric Analysis of Panel Data*. Wiley.

[2] 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. *J. Amer. Statist. Assoc.*, 28-36.

[3] Box, G.E.P. and Cox, D.R. (1964). An analysis of transformation (with discussion). *J. Royal Statist. Soc.*, **B26**, 211-252.

[4] Chatterjee, S., Lahiri, P. and Li, H. (2008). Parametric bootstrap approximation to the distribution of EBLUP and related predictions intervals in linear mixed models. *Ann. Statist.*, **36**, 1221-1245.

[5] Datta, G.S., Rao, J.N.K. and Smith, D.D. (2005). On measuring the variability of small area estimators under a basic area level model. *Biometrika*, **92**, 183-196.

[6] Gelfand, A. E., Hills, S. E., Racine-Poon, A. and Smith, A. F. M. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. *J. Amer. Statist. Assoc.*, **85**, 972-985.

[7] Gurka, M., Edwards, L. J., Muller, K. E. and Kupper, L. L. (2006). Extending the Box-Cox transformation to the linear mixed model. *J. Royal Statist. Soc.*, **A169**, 273-288.

[8] Jiang, J. (2010). *Large Sample Techniques for Statistics*. Springer.

[9] Jiang, J. and Lahiri, P. (2006). Mixed model prediction and small area estimation (with discussions). *Test*, **15**, 1-96.

[10] Kawakubo, Y. and Kubokawa, T. (2014). Modified conditional AIC in linear mixed models. *J. Multivariate Anal.*, **129**, 44-56.

[11] Lee, J. C., Lin, T. I., Lee, K. J. and Hsu, Y. L. (2005). Bayesian analysis of Box-Cox transformed linear mixed models with ARMA(p, q) dependence. *J. Statist. Plan. Infer.*, **133**, 435-451.

[12] McCulloch, C. E. and Searle, S. R. (2001). *Generalized, Linear, and Mixed Models*. Wiley, New York.

[13] Rao, J.N.K. (2003). *Small Area Estimation*. Wiley, New York.

[14] Sugasawa, S. and Kubokawa, T. (2014a). Parametric transformed Fay-Herriot model for small area estimation. *ArXiv:1410.4778.*

[15] Sugasawa, S. and Kubokawa, T. (2014b). Estimation and prediction in transformed nested error regression model. *ArXiv:1410.8269.*

[16] Verbeke, G. and Molenberghs, G. (2009). *Linear Mixed Models for Longitudinal Data*. Springer.

[17] Yang, Z. L. (2006). A modified family of power transformations. *Econ. Letters*, **92**, 14-19.