CIRJE-F-929

Estimation and Prediction Intervals in Transformed Linear Mixed Models

Shonosuke Sugasawa
Graduate School of Economics, The University of Tokyo

Tatsuya Kubokawa
The University of Tokyo

April 2014

Estimation and Prediction Intervals in Transformed Linear Mixed Models

Shonosuke Sugasawa* and Tatsuya Kubokawa†

University of Tokyo

April 17, 2014

Abstract

For analyzing positive or bounded data, this paper suggests parametrically transformed nested error regression models (TNERM), which not only include the log-transformed model, but also adjust flexibly the transformation parameter to fit the data to a normal linear regression. Conditions on the transformation are derived for consistency of the maximum likelihood estimator for the transformation parameter. The conditions are satisfied by the dual power transformation for positive data and the dual power logistic transformation for bounded data. In order to calibrate uncertainty of the transformed empirical best linear unbiased predictor (TEBLUP), the paper derives prediction intervals with second-order accuracy based on the parametric bootstrap method. Conditional prediction intervals given data in the area of interest are also constructed. The proposed methods are investigated through simulation and empirical studies.

Key words and phrases: Box-Cox transformation, dual power transformation, linear mixed model, nested error regression model, parametric bootstrap, prediction intervals, small area estimation.

1 Introduction

The linear mixed models with both random and fixed effects have been extensively and actively studied in recent years from both theoretical and applied aspects in the literature. Of these, the nested error regression model (NERM) has been used as a unit level linear mixed model in the framework of small-area estimation. Since direct estimates like sample means for small areas have unacceptable estimation errors because of small sample sizes in small areas, the model-based shrinkage methods such as the empirical best linear unbiased predictor (EBLUP) are very useful for providing reliable estimates for small-areas with higher precisions by borrowing

*Graduate School of Economics, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, JAPAN, E-Mail: shonosuke622@gmail.com

†Faculty of Economics, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, JAPAN, E-Mail: tatsuya@e.u-tokyo.ac.jp

data in the surrounding areas. For a good survey on this topic, see Ghosh and Rao (1994) and Rao (2003).

A convincing example of NERM and EBLUP was given by Battese, Harter and Fuller (1988), who analyzed data of crop areas in m counties for $m = 12$. From the i-th county, n_i segments are sampled. Each segment is about 250 hectares, and the area of corn (or soybeans) in the j-th segment, denoted by y_{ij}, is reported as survey data by interviewing farm operators. For the j-th segment, on the other hand, the numbers of pixels (0.45 hectares) classified as corn and soybeans, denoted by x_{1ij} and x_{2ij}, are available from LANDSAT satellite data. Battese, *et al.* (1988) analyzed the data successfully using the nested error regression model (NERM) in the framework of a finite population. Without assuming the finite population model, NERM is described as

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + v_i + \varepsilon_{ij}, \tag{1}$$

and the problem is the prediction of the mean of crop areas in the i-th county, denoted by

$$\xi_i = \beta_0 + \beta_1 \overline{x}_{1i} + \beta_2 \overline{x}_{2i} + v_i,$$

for $\overline{x}_{1i} = \sum_{j=1}^{n_i} x_{1ij}/n_i$ and $\overline{x}_{2i} = \sum_{j=1}^{n_i} x_{2ij}/n_i$, where v_i's and ε_{ij}'s are mutually independently distributed as $v_i \sim \mathcal{N}(0, \sigma_v^2)$ and $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_e^2)$. For analyzing such unit level data, NERM and its variants in finite population models are useful. When applying data to NERM, the following queries are raised:

(I) We are often faced with the case that y_{ij}'s are positive values like the crop areas data. Fitting such positive data to the normal distributions in NERM may be inappropriate in the case that the distribution of y_{ij}'s is skewed. As investigated in Slud and Maiti (2006), an alternate method is to apply the log-transformed data to NERM. However, it depends on a feature of data whether we treat the original observations y_{ij} or the log-transformed ones $\log(y_{ij})$. Which method should we use for given data?

(II) As explained above, the crop areas data y_{ij}'s are bounded above from 250 hectares. For such bounded data, how should we transform the data to fit to NERM?

A conventional method for the query (I) is the Box-Cox transformation suggested by Box and Cox (1964), described by

$$h^{BC}(y_{ij}, \lambda) = \begin{cases} (y_{ij}^{\lambda} - 1)/\lambda, & \lambda \neq 0, \\ \log y_{ij}, & \lambda = 0. \end{cases}$$

However, it is known that the the maximum likelihood (ML) estimator of the transformation parameter λ in the Box-Cox transformation is not consistent, which means that EBLUP is not consistent with BLUP asymptotically. An alternative is the dual power transformation suggested by Yang (2006), which is given by

$$h^{DP}(y_{ij}, \lambda) = \begin{cases} (y_{ij}^{\lambda} - y_{ij}^{-\lambda})/2\lambda, & \lambda > 0, \\ \log y_{ij}, & \lambda = 0. \end{cases}$$

This is a transformation from positive numbers to real numbers, and the ML estimator of λ is consistent. For the query (II), we suggest the dual power logistic transformation (DPLT). When y_{ij} is restricted on the interval $(0, 1)$, the DPLT is given by

$$h^{DPL}(y_{ij}, \lambda) = \begin{cases} \left\{ \left(\frac{y_{ij}}{1-y_{ij}}\right)^{\lambda} - \left(\frac{1-y_{ij}}{y_{ij}}\right)^{\lambda} \right\} / 2\lambda, & \lambda > 0, \\ \log\left(\frac{y_{ij}}{1-y_{ij}}\right), & \lambda = 0. \end{cases}$$

In this paper, we suggest a general class of parametric transformations $h(y_{ij}, \lambda)$ and consider the parametrically transformed NERM (TNERM) described as $h(y_{ij}, \lambda) = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + v_i + \varepsilon_{ij}$. This model has the transformation parameter λ which can be used for adjustment, so that TNERM enables us to flexibly analyze the small-area positive or bounded data. We clarify conditions on $h(\cdot, \lambda)$ under which the ML estimator $\widehat{\lambda}$ of λ is consistent. It is noted that those conditions are satisfied by both transformations $h^{DP}(y_{ij}, \lambda)$ and $h^{DPL}(y_{ij}, \lambda)$. Base on the consistency of $\widehat{\lambda}$, we suggests consistent estimators for the parameters $\boldsymbol{\beta}$, σ_v^2 and σ_e^2.

For the prediction of ξ_i, the consistent estimators are substituted into BLUP to get EBLUP, denoted by $\hat{\xi}_i^{EB}$, which is consistent with BLUP asymptotically. However, an interesting quantity is the inversely transformed value of ξ_i, namely,

$$h^{-1}(\xi_i, \lambda)$$

rather than ξ_i. Then, the predictor induced from the EBLUP $\hat{\xi}_i^{EB}$ is $h^{-1}(\hat{\xi}_i^{EB}, \widehat{\lambda})$, which we call the transformed EBLUP (TEBLUP). Since TEBLUP is expected to give reliable predicted values for small-areas with higher precisions, it is important to assess uncertainty of TEBLUP, and the following query is raised:

(III) How can we measure uncertainty of TEBLUP $h^{-1}(\hat{\xi}_i^{EB}, \widehat{\lambda})$ as a predictor of $h^{-1}(\xi_i, \lambda)$? An approach to measuring the uncertainty is to provide an estimate of the mean squared error (MSE) of TEBLUP. Sugasawa and Kubokawa (2013) gave second-order unbiased estimators of MSE of $\hat{\xi}_i^{EB}$ in the parametrically transformed Fay-Herriot model. However, it is not necessarily appropriate to measure uncertainty of TEBLUP $h^{-1}(\hat{\xi}_i^{EB}, \widehat{\lambda})$ with MSE as well as a second-order unbiased estimator of MSE for TEBLUP is hard to derive.

For the query (III), in this paper, we consider to construct a prediction interval of $h^{-1}(\xi_i, \lambda)$ based on $h^{-1}(\hat{\xi}_i^{EB}, \widehat{\lambda})$. Since it is harder to derive an analytical prediction interval with suitable accuracy based on the Taylor series expansion, we here provide a prediction interval with second-order accuracy based on the parametric bootstrap along the line given in Chatterjee, Lahiri and Li (2008). We also provide a conditional prediction interval given data in the area of interest. This corresponds to the results of Booth and Hobert (1998) who discussed a conditional MSE and its estimation.

The paper is organized as follows: In Section 2, we suggest the parametric transformed nested error regression model (TNERM) for a general class of transformations. We also provide conditions on transformations which guarantee consistency of the ML estimator for the transformation parameter. As useful transformations, we treat the dual power transformation (DPT) proposed by Yang (2006) and the dual power logistic transformation (DPLT), which is newly proposed based on motivation from DPT and logistic transformation. Some consistent estimators of parameters in TNERM are also given. In Section 3, we introduce the transformed EBLUP (TEBLUP) and construct unconditional and conditional prediction intervals based on TEBLUP using the parametric bootstrap method. It is shown that these prediction intervals have second-order accuracy for nominal confidence coefficient $1-\alpha$. In Section 4, a finite-sample performance of unconditional and conditional prediction intervals of TEBLUP through simulation. As an empirical study, in Section 5, we revisit the data given in Battese, *et al.* (1988) who handled the original observations y_{ij}'s. Applying the DP transformation to the data, we observe that the transformation parameter λ is estimated by zero, and this suggests that the log-transformed model should be used for DPT. Since the crop areas data are bounded above from 250 hectares, we can also apply the DP logistic transformation. Unconditional and conditional prediction intervals are given for the log-transformed NERM and the DP logistic transformed NERM. The concluding remarks are given in Section 6, and the technical proofs are given in the Appendix.

2 Parametric Transformations of the Nested Error Regression Model (NERM)

2.1 Transformed NERM

Consider the two-stage cluster sampling, namely, m clusters are randomly selected, and data are randomly selected from each selected cluster. For $i=1,\ldots,m$, a random sample taken from the i-th cluster with size n_i is denoted by $y_{i1},\ldots,y_{in_i}$. The most useful model for analyzing such data is the nested error regression model (NERM) described by

$$y_{ij}=\boldsymbol{x}_{ij}'\boldsymbol{\beta}+v_i+\varepsilon_{ij},\quad i=1,\ldots,m,\ j=1,\ldots,n_i, \tag{2}$$

where v_i's and ε_{ij}'s are mutually independently distributed as $v_i\sim\mathcal{N}(0,\sigma_v^2)$ and $\varepsilon_{ij}\sim\mathcal{N}(0,\sigma_e^2)$. Here, a vector $\boldsymbol{x}'$ denotes the transpose of $\boldsymbol{x}$, $\mathcal{N}(\mu,\sigma^2)$ denotes a normal distribution with mean μ and variance σ^2, $\boldsymbol{x}_{ij}$ is a p-dimensional known covariate associated with y_{ij}, $\boldsymbol{\beta}$ is a p-dimensional unknown vector of regression coefficients, and σ_v^2 and σ_e^2 are unknown components of variance, called 'between' and 'within' components, respectively.

The model (2) is a linear mixed model which incorporates both fixed and random effects, and it has been used for analyzing unit level data in the framework of small-area estimation. If y_{ij}'s are real-valued data, the model (2) is reasonable. However, it is inappropriate when values of y_{ij}'s are limited to $\mathbb{R}_+$ or $\mathbb{R}_{(0,1)}$, where $\mathbb{R}_+=\{x\in\mathbb{R}; x>0\}$ and $\mathbb{R}_{(a,b)}=\{x\in\mathbb{R}; a<x<b\}$ for the real space $\mathbb{R}$. In this paper, it is assumed that the data take values in D which is a subset of $\mathbb{R}$, namely, $y_{ij}\in D$. For example, $D=\mathbb{R}_+$ if y_{ij} is positive, and $D=\mathbb{R}_{(0,1)}$ if $y_{ij}\in(0,1)\subset\mathbb{R}$. Then, we need to consider a transformation of $y_{ij}\in D$ to fit NERM. Let $h(\cdot,\lambda)$ be a monotone increasing transformation from D to $\mathbb{R}$ for given λ. The parameter λ is adjusted so that transformed data $h(y_{ij},\lambda)$'s can fit NERM. Thus, we can suggest the parametrically transformed nested error regression model (TNERM)

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

It may be convenient to write the model (3) in matricial forms. Let $\boldsymbol{y} = (y_{i1}, \dots, y_{in_i})'$, $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}, \dots, \boldsymbol{x}_{in_i})'$, $\boldsymbol{\epsilon}_i = (\varepsilon_{i1}, \dots, \varepsilon_{in_i})'$ and $\boldsymbol{j}_{n_i} = (1, \dots, 1) \in \mathbb{R}^{n_i}$. Also, define $\boldsymbol{h}(\boldsymbol{y}_i, \lambda)$ by

$$\boldsymbol{h}(\boldsymbol{y}_i, \lambda) = (h(y_{i1}, \lambda), \dots, h(y_{in_i}, \lambda))'.$$

Then, the model (3) is expressed as

$$\boldsymbol{h}(\boldsymbol{y}_i, \lambda) = \boldsymbol{X}_i \boldsymbol{\beta} + \boldsymbol{j}_{n_i} v_i + \boldsymbol{\epsilon}_i, \quad i = 1, \dots, m, \tag{4}$$

and $\boldsymbol{h}(\boldsymbol{y}_i, \lambda)$ has an n_i-variate normal distribution $\mathcal{N}_{n_i}(\boldsymbol{X}_i \boldsymbol{\beta}, \sigma_e^2 \boldsymbol{V}_i(\rho))$ where $\boldsymbol{V}_i(\rho) = \boldsymbol{I}_{n_i} + \rho \boldsymbol{J}_{n_i}$ for $\rho = \sigma_v^2/\sigma_e^2$, the $n_i \times n_i$ identity matrix $\boldsymbol{I}_{n_i}$ and $\boldsymbol{J}_{n_i} = \boldsymbol{j}_{n_i} \boldsymbol{j}'_{n_i}$. It is noted that the covariance of $h(\boldsymbol{y}_i, \lambda)$ has the intra-class correlation structure, namely $h(y_{i1}, \lambda), \dots, h(y_{in_i}, \lambda)$ are not mutually independent when $\rho \neq 0$. Let $N = \sum_{i=1}^m n_i$. All the data $\boldsymbol{y}_i$'s are described as the N-dimensional vector $\boldsymbol{Y} = (\boldsymbol{y}'_i, \dots, \boldsymbol{y}'_m)'$. Then the joint density function $\boldsymbol{Y}$ is expressed as

$$\begin{aligned} f(\boldsymbol{Y}) = {} & (2\pi)^{-N/2} \sigma_e^{-m} \prod_{i=1}^m \det(\boldsymbol{V}_i(\rho)) \prod_{i=1}^m \prod_{j=1}^{n_i} h_x(y_{ij}, \lambda) \\ & \times \exp\Big\{ -\frac{1}{2} \sigma_e^{-2} \sum_{i=1}^m (\boldsymbol{h}(\boldsymbol{y}_i, \lambda) - \boldsymbol{X}_i \boldsymbol{\beta})' \boldsymbol{V}_i(\rho)^{-1} (\boldsymbol{h}(\boldsymbol{y}_i, \lambda) - \boldsymbol{X}_i \boldsymbol{\beta}) \Big\}, \end{aligned} \tag{5}$$

where $\prod_{i=1}^m \prod_{j=1}^{n_i} h_x(y_{ij}, \lambda)$ is the Jacobian of the transformation for $h_x(x, \lambda) = \partial h(x, \lambda)/\partial x$. This expression will be used for estimating the unknown parameters $\boldsymbol{\beta}$, σ_v^2, σ_e^2 and λ.

For the parametric transformation $h(x, \lambda)$ given in the model (3), we need to assume conditions under which existence of an estimator of λ and its consistency are guaranteed. For notational convenience, let $h_{a_1 a_2 \cdots a_n}(x, \lambda)$ for $a_1, a_2, \dots a_n \in \{x, \lambda\}$ be the partial derivative of $h(x, \lambda)$, i.e. $h_{a_1 a_2 \cdots a_n}(x, \lambda) = \partial^n f(x, \lambda)/\partial a_1 \dots \partial a_n$. For example, $h_{xx}(x, \lambda) = \partial^2 h(x, \lambda)/\partial x^2$, $h_{x\lambda}(x, \lambda) = \partial^2 h(x, \lambda)/\partial x \partial \lambda$ and others. Moreover $h_{a_1 a_2 \cdots a_n}(c, \lambda)$ or $h_{a_1 a_2 \cdots a_n}(x, c)$ for $c \in \mathbb{R}$ means that $h_{a_1 a_2 \cdots a_n}(x, \lambda)|_{x=c}$ or $h_{a_1 a_2 \cdots a_n}(x, \lambda)|_{\lambda=c}$ respectively.

Assumption 1. The function $h(\cdot, \lambda)$ is a transformation from D to $\mathbb{R}$ which is characterized by the parameter $\lambda \in \Lambda (\subset \mathbb{R})$ and satisfies the following:

(A.1) $h(x, \lambda)$ is a monotone increasing function of x ($x \in D$) and its range is $\mathbb{R}$.

(A.2) $h(x, \lambda)$ and $h^{-1}(x, \lambda)$ are three times continuously differentiable, where $f(x, \lambda) = h^{-1}(x, \lambda)$ is the inverse function of $h(x, \lambda)$ defined by $x = h(f(x, \lambda), \lambda)$.

(A.3) The moments of the following exist for each fixed $\lambda \in \Lambda$:
(1) $\{h_\lambda(x, \lambda)\}^2$ and $\{(\partial/\partial\lambda) \log(h_x(x, \lambda))\}^2$,
(2) $\{h_\lambda(x, \lambda)\}^4$, $\{h_{\lambda\lambda}(x, \lambda)\}^2$ and $\{(\partial^2/\partial\lambda^2) \log(h_x(x, \lambda))\}^2$,
(3) $h^{-1}(x, \lambda)$, $h_x^{-1}(x, \lambda)$, $h_\lambda^{-1}(x, \lambda)$, $h_{\lambda x}^{-1}(x, \lambda)$, $h_{xx}^{-1}(x, \lambda)$ and $h_{\lambda\lambda}^{-1}(x, \lambda)$,
where their expectations are taken with respect to $h(x, \lambda)$ which is normally distributed.

Condition (A.1) means that the transformation is a one-to-one and onto function from D to $\mathbb{R}$. Clearly, (A.1) is not satisfied by the Box-Cox transformation (Box and Cox, 1964), but by the logarithmic transformation. Conditions (A.2) and (A.3) will be used for establishing consistency of estimators including transformation parameter λ and for constructing prediction intervals. Especially, (A.2) and (A.3) (1) guarantees that the random variable $F(\widehat{\boldsymbol{\beta}}(\lambda), \widehat{\sigma}_v^2(\lambda), \widehat{\sigma}_e^2(\lambda), \lambda)$ given in (12) converges in probability, and (A.3)(2) guarantees that $(\partial/\partial\lambda)F(\widehat{\boldsymbol{\beta}}(\lambda), \widehat{\sigma}_v^2(\lambda), \widehat{\sigma}_e^2(\lambda), \lambda)$ converges in probability.

We here provide a couple of examples of the transformations which satisfy Assumption 1. One is the dual power transformation suggested by Yang (2006) for analyzing positive data, the other is the dual power logistic transformation which is useful for analyzing data taken in the interval $D = (0, 1)$. The dual power logistic transformation has not been known as long as we know.

Example 1 (Dual Power Transformations)**.** For $x > 0$, the dual power transformation (DPT) is described as

$$h^{DP}(x, \lambda) = \begin{cases} (x^{\lambda} - x^{-\lambda})/2\lambda, & \lambda > 0, \\ \log x, & \lambda = 0. \end{cases} \tag{6}$$

Although the Box-Cox transformation does not satisfy assumption (A.1), the dual power transformation satisfies (A.1). As shown later, the maximum likelihood estimator of λ for $h^{DP}(x, \lambda)$ is consistent, while the MLE of the transformation parameter in the Box-Cox transformation is not consistent. This shows that the dual power transformation is useful for constructing prediction intervals by replacing λ with the MLE. It is noted that for $z = h^{DP}(x, \lambda)$, the inverse transformation is expressed as

$$x = \left(\lambda z + \sqrt{\lambda^2 z^2 + 1}\right)^{1/\lambda}$$

for $\lambda \neq 0$, and $x = e^z$ for $\lambda = 0$. Then, it can be easily shown that DPT satisfies Assumption 1.

When data are restricted on the space $\{x \in \mathbb{R} | x > a\}$ for $a \in \mathbb{R}$, DPT can be extended to $h^{DP}(x - a, \lambda)$ for analyzing data on the space. □

Example 2 (Dual Power Logistic Transformation)**.** To analyze the data in the interval $D = (0, 1)$, we newly propose the dual power logistic transformation (DPLT). This is naturally induced from DPT by replacing x in $h^{DP}(x, \lambda)$ with the odd $(1 - x)/x$. Thus, for $0 < x < 1$, the DPLT is given by

$$h^{DPL}(x, \lambda) = \begin{cases} \left\{\left(\frac{x}{1-x}\right)^{\lambda} - \left(\frac{1-x}{x}\right)^{\lambda}\right\}/2\lambda, & \lambda > 0, \\ \log\left(\frac{x}{1-x}\right), & \lambda = 0. \end{cases} \tag{7}$$

Using the expression of the inverse transformation of DPT, one gets the inverse transformation of DPLT, given by

$$x = \frac{\left(\lambda z + \sqrt{\lambda^2 z^2 + 1}\right)^{1/\lambda}}{1 + \left(\lambda z + \sqrt{\lambda^2 z^2 + 1}\right)^{1/\lambda}}$$

for $\lambda \neq 0$, and $x = e^z/(1 + e^z)$ for $\lambda = 0$. Then, it can be easily shown that DPLT satisfies Assumption 1.

When data are restricted on the interval (a, b) for fixed values a and b, $(a < b)$, DPLT can be extended to $h^{DPL}((x-a)/(b-a), \lambda)$, since $(x-a)/(b-a)$ lies in $(0, 1)$. Thus, we can analyze data on (a, b) using $h^{DPL}((x - a)/(b - a), \lambda)$. □

2.2 Consistent estimators of the parameters

We here provide consistent estimators of the unknown parameters $\boldsymbol{\beta}$, σ_v^2, σ_e^2 and λ. To this end, we begin by estimating $\boldsymbol{\beta}$, σ_v^2 and σ_e^2 in the case of known λ. In this case, the conventional estimators given in the literature for NERM (2) can be inherited to the transformed model.

Concerning estimation of $\boldsymbol{\beta}$, the maximum likelihood (ML) or generalized least square (GLS) estimator of $\boldsymbol{\beta}$ for known σ_v^2, σ_e^2 and λ is

$$\widehat{\boldsymbol{\beta}}(\rho, \lambda) = \Big(\sum_{i=1}^m \frac{n_i \bar{\boldsymbol{x}}_i \bar{\boldsymbol{x}}_i'}{1 + n_i \rho}\Big)^{-1} \sum_{i=1}^m \frac{n_i \bar{\boldsymbol{x}}_i z_i(\lambda)}{1 + n_i \rho}, \tag{8}$$

where $\rho = \sigma_v^2/\sigma_e^2$, $\bar{\boldsymbol{x}}_i = \sum_{j=1}^{n_i} \boldsymbol{x}_{ij}/n_i$ is the mean of covariates $\boldsymbol{x}_{ij}$'s for the i-th area, and

$$z_i(\lambda) = \frac{1}{n_i} \sum_{j=1}^{n_i} h(y_{ij}, \lambda), \qquad i = 1, \ldots, m,$$

is the mean of the transformed observations. Since $\widehat{\boldsymbol{\beta}}(\rho, \lambda) \sim \mathcal{N}_p(\boldsymbol{\beta}, \{\sum_{j=1}^m (n_j \sigma_v^2 + \sigma_e^2)^{-1} n_j \bar{\boldsymbol{x}}_j \bar{\boldsymbol{x}}'\}^{-1})$, it is clear that $\widehat{\boldsymbol{\beta}}(\rho, \lambda)$ is consistent and $\widehat{\boldsymbol{\beta}}(\rho, \lambda) - \boldsymbol{\beta} = \boldsymbol{O}_p(m^{-1/2})$ under the following assumption:

Assumption 2. *The following are assumed for $\bar{\boldsymbol{x}}_i$ and n_i:*

(A.4) $m^{-1} \sum_{i=1}^m \bar{\boldsymbol{x}}_i \bar{\boldsymbol{x}}_i'$ converges to a positive definite matrix.

(A.5) There exist integers $\underline{n}$ and $\overline{n}$ which are positive and independent of m such that $\underline{n} \leq n_i \leq \overline{n}$ for $i = 1, \ldots, m$.

Since σ_v^2 and σ_e^2 are unknown, we estimate them and then substitute their estimators into $\widehat{\boldsymbol{\beta}}(\rho, \lambda)$. In NERM (2) with known λ, for σ_v^2 and σ_e^2, the Prasad-Rao estimator, the maximum likelihood (ML) and the restricted maximum likelihood (REML) estimators have been used in the literature, and it would be plausible that those estimators can be used still in TNERM (3) by replacing λ with an estimator. We here clarify conditions that estimators of σ_v^2 and σ_e^2 should satisfy in order to derive prediction intervals given in this paper. For notational convenience, $\boldsymbol{O}_p(a_n)$ means that every component in $\boldsymbol{O}_p(a_n)$ is of order $O_p(a_n)$, and the notation $\boldsymbol{O}(a_n)$ is defined similarly.

Assumption 3. Let $\widehat{\boldsymbol{\sigma}}^2(\lambda) = (\widehat{\sigma}_e^2(\lambda), \widehat{\sigma}_v^2(\lambda))'$ be an estimator of $\boldsymbol{\sigma}^2 = (\sigma_e^2, \sigma_v^2)'$ in the case of known λ. Then it is assumed that the estimator $\widehat{\boldsymbol{\sigma}}^2(\lambda)$ satisfies the following:

(A.6) $(\widehat{\boldsymbol{\sigma}}^2(\lambda) - \boldsymbol{\sigma}^2)|\boldsymbol{y}_i = \boldsymbol{O}_p(m^{-1/2})$.

(A.7) $E[\widehat{\boldsymbol{\sigma}}^2(\lambda) - \boldsymbol{\sigma}^2|\boldsymbol{y}_i] = \boldsymbol{O}_p(m^{-1})$.

(A.8) $\partial\widehat{\boldsymbol{\sigma}}^2(\lambda)/\partial\lambda\big|\boldsymbol{y}_i = \boldsymbol{O}_p(1)$.

(A.9) $\Big(\partial\widehat{\boldsymbol{\sigma}}^2(\lambda)/\partial\lambda - E\big[\partial\widehat{\boldsymbol{\sigma}}^2(\lambda)/\partial\lambda\big|\boldsymbol{y}_i\big]\Big)\big|\boldsymbol{y}_i = \boldsymbol{O}_p(m^{-1/2})$.

Condition (A.6) implies that the estimators $\widehat{\sigma}_v^2(\lambda)$ and $\widehat{\sigma}_e^2(\lambda)$ are consistent. Conditions (A.7) and (A.8) will be used for investigating asymptotic properties of $\widehat{\boldsymbol{\sigma}}^2(\widehat{\lambda})$.

Substituting $\widehat{\rho}(\lambda) = \widehat{\sigma}_v^2(\lambda)/\widehat{\sigma}_e^2(\lambda)$ into $\widehat{\boldsymbol{\beta}}(\rho, \lambda)$, one gets the estimator $\widehat{\boldsymbol{\beta}}(\lambda)$ defined by

$$\widehat{\boldsymbol{\beta}}(\lambda) = \widehat{\boldsymbol{\beta}}(\widehat{\rho}(\lambda), \lambda).$$

It is noted from (A.8) that $\widehat{\rho}(\lambda) - \rho = O_p(m^{-1/2})$. Some asymptotic properties on $\widehat{\boldsymbol{\beta}}(\lambda)$ are given in the following lemma which will be proved in the Appendix. Lemma 1 will be used in Theorem 1 for showing the second-order accuracy of the parametric bootstrap procedure.

Lemma 1 (Asymptotic properties of $\widehat{\boldsymbol{\beta}}(\lambda)$)**.** *Under Assumptions 1, 2 and 3, it holds that* $(\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta})\big|\boldsymbol{y}_i = \boldsymbol{O}_p(m^{-1/2})$, $E[\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}|\boldsymbol{y}_i] = \boldsymbol{O}_p(m^{-1})$ *and*

$$\Big(\partial\widehat{\boldsymbol{\beta}}(\lambda)/\partial\lambda - E[\partial\widehat{\boldsymbol{\beta}}(\lambda)/\partial\lambda\big|\boldsymbol{y}_i]\Big)\Big|\boldsymbol{y}_i = \boldsymbol{O}_p(m^{-1/2}).$$

We here provide some conventional estimators of σ_v^2 and σ_e^2 and show whether those estimators satisfy Assumption 3.

[1] Prasad-Rao estimator. Let $\boldsymbol{X} = (\boldsymbol{X}_1', \ldots, \boldsymbol{X}_m')'$ and $\boldsymbol{E} = \text{blockdiag}(\boldsymbol{E}_1, \ldots, \boldsymbol{E}_m)$ for $\boldsymbol{E}_i = \boldsymbol{I}_{n_i} - n_i^{-1}\boldsymbol{J}_i$. Defined $\boldsymbol{h}(\boldsymbol{Y}, \lambda)$ by

$$\boldsymbol{h}(\boldsymbol{Y}, \lambda) = (\boldsymbol{h}(\boldsymbol{y}_1, \lambda)', \ldots, \boldsymbol{h}(\boldsymbol{y}_m, \lambda)')'.$$

Then define S_1 and S_2 by $S_1 = \boldsymbol{h}(\boldsymbol{Y}, \lambda)'(\boldsymbol{I}_N - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{h}(\boldsymbol{Y}, \lambda)$ and $S_2 = \boldsymbol{h}(\boldsymbol{Y}, \lambda)'(\boldsymbol{E} - \boldsymbol{E}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{E}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{E})\boldsymbol{h}(\boldsymbol{Y}, \lambda)$. Then, Prasad and Rao (1990) suggested unbiased estimators of σ_v^2 and σ_e^2 given by

$$\widehat{\sigma}_{e.PR}^2 = \frac{S_2}{N - m - p} \qquad \text{and} \qquad \widehat{\sigma}_{v.PR}^2 = \frac{1}{N_*}\big\{S_1 - (N - p)\widehat{\sigma}_{e.PR}^2\big\}, \tag{9}$$

where $N = \sum_{i=1}^m n_i$ and $N_* = N - \text{tr}\,\{(\boldsymbol{X}'\boldsymbol{X})^{-1}\sum_{i=1}^m n_i^2\bar{\boldsymbol{x}}_i'\bar{\boldsymbol{x}}_i\}$. The Prasad–Rao estimator estimator of $\boldsymbol{\sigma}^2$ is denoted by $\widehat{\boldsymbol{\sigma}}_{PR}^2 = (\widehat{\sigma}_{e.PR}^2, \widehat{\sigma}_{v.PR}^2)'$. It is noted that $N = O(m)$, $N - m - p = O(m)$ and $N^* = O(m)$ under Assumption 2.

[2] ML. The maximum likelihood (ML) estimator $\widehat{\boldsymbol{\sigma}}_{ML}^2 = (\widehat{\sigma}_{e.ML}^2, \widehat{\sigma}_{v.ML}^2)'$ of $\boldsymbol{\sigma}^2 = (\sigma_v^2, \sigma_e^2)'$ are given as the solutions of the equations

$$L_1(\widehat{\boldsymbol{\sigma}}_{ML}^2) = 0 \quad \text{and} \quad L_2(\widehat{\boldsymbol{\sigma}}_{ML}^2) = 0, \tag{10}$$

where

$$L_1(\boldsymbol{\sigma}^2) = \frac{1}{\sigma_e^4}\sum_{i=1}^m \|\boldsymbol{h}(\boldsymbol{y}_i, \lambda) - \boldsymbol{X}_i\widehat{\boldsymbol{\beta}}(\rho, \lambda) - \frac{n_i\rho}{1 + n_i\rho}(z_i(\lambda) - \bar{\boldsymbol{x}}_i'\widehat{\boldsymbol{\beta}}(\rho, \lambda))\boldsymbol{j}_i\|^2 - \sum_{i=1}^m \frac{n_i}{\sigma_e^2}\Big(1 - \frac{\rho}{1 + n_i\rho}\Big),$$

$$L_2(\boldsymbol{\sigma}^2) = \sum_{i=1}^m \frac{n_i^2}{(\sigma_e^2 + n_i\sigma_v^2)^2}\big\{\boldsymbol{z}_i(\lambda) - \bar{\boldsymbol{x}}_i'\widehat{\boldsymbol{\beta}}(\rho, \lambda)\big\}^2 - \sum_{i=1}^m \frac{n_i}{\sigma_e^2 + n_i\sigma_v^2}.$$

[3] REML. The restricted maximum likelihood (REML) estimator $\widehat{\boldsymbol{\sigma}}^2_{RML} = (\widehat{\sigma}^2_{e.RML}, \widehat{\sigma}^2_{v.RML})'$ of $\boldsymbol{\sigma}^2$ is given as the solutions of the equations

$$\begin{aligned} 0 &= L_1(\boldsymbol{\sigma}^2) + \mathrm{tr}\,[(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-2}\boldsymbol{X}], \\ 0 &= L_2(\boldsymbol{\sigma}^2) + \mathrm{tr}\,[(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Z}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}], \end{aligned} \tag{11}$$

where $\boldsymbol{\Sigma} = \mathrm{blockdiag}(\sigma_e^2\boldsymbol{V}_1(\rho), \ldots, \sigma_e^2\boldsymbol{V}_m(\rho))$, the covariance matrix of $\boldsymbol{h}(\boldsymbol{Y}, \lambda)$.

The following lemma guarantees that the above three estimators satisfy Assumption 3, where the proof will be given in the Appendix.

Lemma 2. *Under Assumption 1 and 2, the above three estimators* $\widehat{\boldsymbol{\sigma}}^2_{PR}$, $\widehat{\boldsymbol{\sigma}}^2_{ML}$ *and* $\widehat{\boldsymbol{\sigma}}^2_{RML}$ *satisfy Assumption 3.*

Lemma 2 means that Assumption 3 is not so restrictive, because it is satisfied by the three typical estimators.

Finally, we provide an estimator of the transformation parameter λ based on the estimators $\widehat{\boldsymbol{\beta}}(\lambda)$, $\widehat{\sigma}^2_v(\lambda)$ and $\widehat{\sigma}^2_e(\lambda)$. Using the likelihood (5), we suggest the estimator as a solution of the equation

$$F(\widehat{\boldsymbol{\beta}}(\lambda), \widehat{\sigma}^2_v(\lambda), \widehat{\sigma}^2_e(\lambda), \lambda) = 0, \tag{12}$$

where

$$F(\boldsymbol{\beta}, \sigma_v^2, \sigma_e^2, \lambda) = \sigma_e^{-2}\sum_{i=1}^m \Big(\boldsymbol{h}(\boldsymbol{y}_i, \lambda) - \boldsymbol{X}_i\boldsymbol{\beta}\Big)'\boldsymbol{V}_i(\rho)^{-1}\Big(\frac{\partial}{\partial\lambda}\boldsymbol{h}(\boldsymbol{y}_i, \lambda)\Big) + J(\boldsymbol{Y}, \lambda),$$

for

$$J(\boldsymbol{Y}, \lambda) = \sum_{i=1}^m\sum_{j=1}^{n_i}\frac{h_{x\lambda}(y_{ij}, \lambda)}{h_x(y_{ij}, \lambda)}.$$

Lemma 3. *Under Assumptions 1, 2 and 3, the equation* (12) *includes a solution which is consistent to* λ*. This solution is denoted by* $\widehat{\lambda}$*. Then,* $(\widehat{\lambda} - \lambda)|\boldsymbol{y}_i = O_p(m^{-1/2})$ *and* $E(\widehat{\lambda} - \lambda|\boldsymbol{y}_i) = O_p(m^{-1})$ *under .*

It is easy to see that $E(\widehat{\lambda} - \lambda) = O(m^{-1})$ from Lemma 3 since $E(\widehat{\lambda} - \lambda) = E[E(\widehat{\lambda} - \lambda|\boldsymbol{y}_i)]$. Based on the results given in the above lemmas, we can get the following asymptotic properties of estimators of the unknown parameters in TNERM (3). The proof is given in the Appendix.

Lemma 4. *Let* $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma_v^2, \sigma_e^2, \lambda)'$ *and* $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}(\widehat{\lambda})', \sigma_v^2(\widehat{\lambda}), \sigma_e^2(\widehat{\lambda}), \widehat{\lambda})'$*. Under Assumptions 1, 2 and 3, we have* $(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})|\boldsymbol{y}_i = \boldsymbol{O}_p(m^{-1/2})$ *and* $E(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}|\boldsymbol{y}_i) = \boldsymbol{O}_p(m^{-1})$ *for* $i = 1, \ldots, m$*.*

The latter property that $E(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}|\boldsymbol{y}_i) = \boldsymbol{O}_p(m^{-1})$ is technical but crucial for the proof of theorem 1 in Section 3, which gives validity of bootstrap method for constructing prediction intervals of TEBLUP.

3 TEBLUP and Prediction Intervals

We now provide the transformed empirical best linear unbiased predictor (TEBLUP) for small area estimation and construct the prediction intervals based on TEBLUP as a measure of uncertainty of the predictor. Since TEBLUP includes the estimators of the parameters $\boldsymbol{\beta}$, σ_v^2, σ_e^2 and λ, it is difficult to construct an exact prediction interval. Thus, in this section, we try to construct a prediction interval with the second-order accuracy. To this end, the asymptotic results given in the lemmas in the previous section are heavily used.

3.1 TEBLUP

We here consider the problem of predicting the quantity

$$h^{-1}(\xi_i, \lambda) \quad \text{for} \quad \xi_i = \bar{\boldsymbol{x}}_i'\boldsymbol{\beta} + v_i.$$

When $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma_v^2, \sigma_e^2, \lambda)$ is known, it is well known that the conditional distribution of ξ_i given $\boldsymbol{y}_i$ is $\mathcal{N}(\hat{\xi}_i(\boldsymbol{\theta}), \sigma_i^2)$, where

$$\hat{\xi}_i = \hat{\xi}_i(\boldsymbol{\theta}) = E[\xi_i|\boldsymbol{y}_i] = \bar{\boldsymbol{x}}_i'\boldsymbol{\beta} + \frac{n_i\rho}{1+n_i\rho}(z_i(\lambda) - \bar{\boldsymbol{x}}_i'\boldsymbol{\beta}), \tag{13}$$

and

$$\sigma_i^2 = \sigma_i^2(\boldsymbol{\sigma}^2) = \frac{\sigma_v^2}{1+n_i\rho}. \tag{14}$$

The estimator $\hat{\xi}_i(\boldsymbol{\theta})$ is the Bayes estimator of ξ_i in the Bayesian context. Substituting the GLS $\widehat{\boldsymbol{\beta}}(\rho, \lambda)$ given in (8) into (13) yields the predictor

$$\bar{\boldsymbol{x}}_i'\widehat{\boldsymbol{\beta}}(\rho, \lambda) + \frac{n_i\rho}{1+n_i\rho}(z_i(\lambda) - \bar{\boldsymbol{x}}_i'\widehat{\boldsymbol{\beta}}(\rho, \lambda)).$$

It is known that this estimator is the best linear unbiased predictor (BLUP) of ξ_i. For σ_v^2, σ_e^2 and λ, we substitute the estimators given in Section 2 into the BLUP, and the resulting predictor is given by

$$\hat{\xi}_i^{EB} = \hat{\xi}_i(\widehat{\boldsymbol{\theta}}) = \bar{\boldsymbol{x}}_i'\widehat{\boldsymbol{\beta}} + \frac{n_i\widehat{\rho}}{1+n_i\widehat{\rho}}(z_i(\widehat{\lambda}) - \bar{\boldsymbol{x}}_i'\widehat{\boldsymbol{\beta}}), \tag{15}$$

where, for simplicity, we use the notations $\widehat{\boldsymbol{\beta}}$, $\widehat{\sigma}_v^2$, $\widehat{\sigma}_e^2$ and $\widehat{\rho}$ as abbreviation of $\widehat{\boldsymbol{\beta}}(\widehat{\lambda})$, $\widehat{\sigma}_v^2(\widehat{\lambda})$, $\widehat{\sigma}_e^2(\widehat{\lambda})$ $\widehat{\rho}(\widehat{\lambda}) = \widehat{\sigma}_v^2(\widehat{\lambda})/\widehat{\sigma}_e^2(\widehat{\lambda})$ without any confusion. The predictor (15) is called the empirical best linear unbiased predictor (EBLUP). In the Bayesian context, it corresponds to the empirical Bayes estimator of ξ_i.

Since our interest is in the prediction of $h^{-1}(\xi_i, \lambda)$, we need to make the inverse transformation of $\hat{\xi}_i^{EB}$. It should be remarked that the inverse transformation depends on the unknown transformation parameter λ. Hence, the transformed predictor of $h^{-1}(\xi_i, \lambda)$ is given by

$$h^{-1}(\hat{\xi}_i^{EB}, \hat{\lambda}),$$

which is called the transformed empirical best linear unbiased predictor (TEBLUP). For the purpose of measuring uncertainty of TEBLUP $h^{-1}(\hat{\xi}_i^{EB}, \hat{\lambda})$, we construct the prediction interval via parametric bootstrap methods, which will be derived in the following subsections.

3.2 Unconditional prediction interval based on TEBLUP

We construct a prediction interval of $h^{-1}(\xi_i, \lambda)$ with a second-order accuracy for $\xi_i = \boldsymbol{x}_i'\boldsymbol{\beta} + v_i$. Recall that conditionally $\xi_i|\boldsymbol{y}_i \sim \mathcal{N}(\hat{\xi}_i(\boldsymbol{\theta}), \sigma_i^2)$, where $\hat{\xi}_i(\boldsymbol{\theta})$ and σ_i^2 are given in (13) and (14), respectively. This conditional distribution given $\boldsymbol{y}_i$ implies that

$$\sigma_i^{-1}\big\{h(h^{-1}(\xi_i, \lambda), \lambda) - \hat{\xi}_i(\boldsymbol{\theta})\big\}$$

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

Let $\widehat{\sigma}_i^2 = \widehat{\sigma}_v^2/(1 + n_i\widehat{\rho})$. For $\hat{\xi}_i^{EB}$ given in (15), we want to obtain a distribution of

$$T_i = T_i(\xi_i, \lambda, \widehat{\boldsymbol{\theta}}, \boldsymbol{y}_i) = \widehat{\sigma}_i^{-1}\big\{h(h^{-1}(\xi_i, \lambda), \widehat{\lambda}) - \hat{\xi}_i^{EB}\big\}. \tag{16}$$

This distribution is denoted by $\mathcal{L}_m$. If there were constants a_α and b_α such that $P[a_\alpha \leq \widehat{\sigma}_i^{-1}\big\{h(h^{-1}(\xi_i, \lambda), \widehat{\lambda}) - \hat{\xi}_i^{EB}\big\} \leq b_\alpha] = 1 - \alpha$, one would get a $100(1-\alpha)\%$ prediction interval

$$h^{-1}(\xi_i, \lambda) \in \big[h^{-1}(\hat{\xi}_i^{EB} + a_\alpha\widehat{\sigma}_i, \widehat{\lambda}),\ h^{-1}(\hat{\xi}_i^{EB} + b_\alpha\widehat{\sigma}_i, \widehat{\lambda})\big].$$

However, $h(h^{-1}(\xi_i, \lambda), \widehat{\lambda})$ is directly affected by the randomness of $\widehat{\lambda}$, and the distribution $\mathcal{L}_m$ of (16) 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 estimator $\widehat{\boldsymbol{\theta}}$ is consistent from Lemma 4, it can be seen that $\mathcal{L}_m$ converges to the standard normal distribution as m tends to infinity. By approximating a_α and b_α with quantiles of the standard normal distribution, we can construct a prediction interval of $h^{-1}(\xi_i, \lambda)$. However, the accuracy of this prediction interval is of order $O(m^{-1})$, so that such an approximation does not guarantee enough accuracy.

To obtain a prediction interval with accuracy up to $O(m^{-3/2})$, we consider to estimate the distribution $\mathcal{L}_n$ based on the parametric bootstrap method. Let y_{ij}^*'s be a bootstrap sample which is generated as

$$y_{ij}^* = h^{-1}(\boldsymbol{x}_{ij}'\widehat{\boldsymbol{\beta}} + v_i^* + \varepsilon_{ij}^*, \widehat{\lambda}), \quad i = 1, \ldots, m,\ j = 1, \ldots, n_i,$$

where v_i^*'s and ε_{ij}^*'s are mutually independently distributed as $v_i^* \sim \mathcal{N}(0, \widehat{\sigma}_v^2)$ and $\varepsilon_{ij}^* \sim \mathcal{N}(0, \widehat{\sigma}_e^2)$. The estimator $\widehat{\boldsymbol{\theta}}^* = ((\widehat{\boldsymbol{\beta}}^*)', \widehat{\sigma}_v^{2*}, \widehat{\sigma}_e^{2*}, \widehat{\lambda}^*)'$ is calculated from y_{ij}^*'s with the same methods as used to obtain $\widehat{\boldsymbol{\theta}}$. Let $\hat{\xi}_i^{EB*} = \bar{\boldsymbol{x}}_i'\widehat{\boldsymbol{\beta}}^* + (n_i\widehat{\rho}^*/(1 + n_i\widehat{\rho}^*))(z_i^*(\widehat{\lambda}^*) - \bar{\boldsymbol{x}}_i'\widehat{\boldsymbol{\beta}}^*)$ and $\widehat{\sigma}_i^{2*} = \widehat{\sigma}_v^{2*}/(1 + n_i\widehat{\rho}^*)$ for $\widehat{\rho}^* = \widehat{\sigma}_v^{2*}/\widehat{\sigma}_e^{2*}$ and $z_i^*(\widehat{\lambda}^*) = {n_i}^{-1}\sum_{j=1}^{n_i} h(y_{ij}^*, \widehat{\lambda}^*)$. For $\xi_i^* = \bar{\boldsymbol{x}}_i'\widehat{\boldsymbol{\beta}} + v_i^*$, consider the distribution of

$$T_i^* = (\widehat{\sigma}_i^*)^{-1}\big\{h(h^{-1}(\xi_i^*, \widehat{\lambda}), \widehat{\lambda}^*) - \hat{\xi}_i^{EB*}\big\}, \tag{17}$$

which is denoted by $\mathcal{L}_m^*$. As shown in Theorem 1 given below, the distribution $\mathcal{L}_m$ in (16) can be approximated by the bootstrap distribution $\mathcal{L}_m^*$ with accuracy of order $O_p(m^{-3/2})$. Using this approximation, we then proceed to obtain a prediction interval.

Theorem 1. *Under Assumptions 1, 2 and 3, we have*

$$\sup_{q\in\mathbb{R}} |\mathcal{L}_m(q) - \mathcal{L}_m^*(q)| = O_p(m^{-3/2}). \tag{18}$$

The proof of Theorem 1 is given in the Appendix. A direct application of Theorem 1 is the following result on highly accurate prediction intervals.

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

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

where $\mathcal{L}_m^*(\cdot)$ is the distribution function of T_i^*. Then, one gets the prediction interval of $h^{-1}(\xi_i, \lambda)$ given by

$$I_m = \big[h^{-1}(\hat{\xi}_i^{EB} + q_1\widehat{\sigma}_i, \widehat{\lambda}),\ h^{-1}(\hat{\xi}_i^{EB} + q_2\widehat{\sigma}_i, \widehat{\lambda})\big]. \tag{19}$$

Under Assumptions 1, 2 and 3, it holds that

$$P\big(h^{-1}(\xi_i, \lambda) \in I_m\big) = 1 - \alpha + O(m^{-3/2}). \tag{20}$$

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

3.3 Conditional prediction interval

We next construct a conditional prediction interval given data in the area of interest. When data $\boldsymbol{y}_i$ are observed from the i-th area, Booth and Hobert (1998), Datta, Kubokawa, Molina and Rao (2011) treated the conditional MSE of the EBLUP $\hat{\xi}_i^{EB}$ given $\boldsymbol{y}_i$, namely, $E[(\hat{\xi}_i^{EB} - \xi_i)^2 | \boldsymbol{y}_i]$. This conditional MSE measures how much the EBLUP has an estimation error given this data $\boldsymbol{y}_i$, and this conditional approach may be appealing because it conditions on the data in the area of interest. In this subsection, we construct a conditional prediction interval I_m^c given $\boldsymbol{y}_i$ such that

$$P\big(h^{-1}(\xi_i, \lambda) \in I_m^c | \boldsymbol{y}_i\big) = 1 - \alpha + O_p(m^{-3/2}). \tag{21}$$

To this end, we need to approximate the conditional distribution of $T_i = T_i(\xi_i, \lambda, \widehat{\boldsymbol{\theta}}, \boldsymbol{y}_i) = \widehat{\sigma}_i^{-1}\big\{h(h^{-1}(\xi_i, \lambda), \widehat{\lambda}) - \hat{\xi}_i^{EB}\big\}$ given $\boldsymbol{y}_i$. Denote this conditional distribution by $\mathcal{L}_m^c = \mathcal{L}_m^c(\cdot | \boldsymbol{y}_i)$. The difference between the unconditional and conditional prediction intervals is that the unconditional distribution of T_i is considered in (16), while the conditional distribution of T_i given $\boldsymbol{y}_i$ is treated. It is noted that there is a correlation between ξ_i and $\boldsymbol{y}_i$ in (21), namely, the conditional distribution of ξ_i given $\boldsymbol{y}_i$ is $\mathcal{N}(\hat{\xi}_i(\boldsymbol{\theta}), \sigma_i^2)$ for $\hat{\xi}_i(\boldsymbol{\theta})$ and σ_i^2 given in (13) and (14).

Since it is difficult to derive an exact conditional distribution of T_i given $\boldsymbol{y}_i$, we suggest to approximate it via the parametric bootstrap method. A bootstrap sample is generated as

$$y_{kj}^* = h^{-1}(\boldsymbol{x}_{kj}'\widehat{\boldsymbol{\beta}} + v_k^* + \varepsilon_{kj}^*, \widehat{\lambda}), \quad k \neq i,\ k = 1, \ldots, m,\ j = 1, \ldots, n_k,$$

where v_k^*'s and ε_{kj}^*'s are mutually independently distributed as $v_k^* \sim \mathcal{N}(0, \widehat{\sigma}_v^2)$ and $\varepsilon_{kj}^* \sim \mathcal{N}(0, \widehat{\sigma}_e^2)$. Let $\boldsymbol{y}_k^* = (y_{k1}^*, \ldots, y_{kn_k}^*)'$ for $k \neq i$. Noting that $\boldsymbol{y}_i$ is fixed, we can construct the estimator $\widehat{\boldsymbol{\theta}}_{(i)}^* = ((\widehat{\boldsymbol{\beta}}_{(i)}^*)', \widehat{\sigma}_{v(i)}^{2*}, \widehat{\sigma}_{e(i)}^{2*}, \widehat{\lambda}_{(i)}^*)'$ from

$$\boldsymbol{y}_1^*, \ldots, \boldsymbol{y}_{i-1}^*, \boldsymbol{y}_i, \boldsymbol{y}_{i+1}^*, \ldots, \boldsymbol{y}_m^*,$$

with the same technique as used to obtain $\widehat{\boldsymbol{\theta}}$. Let $\hat{\xi}_{(i)}^{EBc*} = \bar{\boldsymbol{x}}_i'\widehat{\boldsymbol{\beta}}_{(i)}^* + (n_i\widehat{\rho}_{(i)}^*/(1+n_i\widehat{\rho}_{(i)}^*))(z_i(\widehat{\lambda}_{(i)}^*) - \bar{\boldsymbol{x}}_i'\widehat{\boldsymbol{\beta}}_{(i)}^*)$ and $\widehat{\sigma}_{(i)}^{2c*} = \widehat{\sigma}_{v(i)}^{2*}/(1+n_i\widehat{\rho}_{(i)}^*)$ for $\widehat{\rho}_{(i)}^* = \widehat{\sigma}_{v(i)}^{2*}/\widehat{\sigma}_{e(i)}^{2*}$ and $z_i(\widehat{\lambda}_{(i)}^*) = n_i{}^{-1}\sum_{j=1}^{n_i} h(y_{ij}, \widehat{\lambda}_{(i)}^*)$. Let ξ_i^{c*} be a random variable having $\mathcal{N}(\hat{\xi}_i^{EB}, \widehat{\sigma}_i^2)$ for $\hat{\xi}_i^{EB} = \hat{\xi}_i(\widehat{\boldsymbol{\theta}})$. Then, for fixed $\boldsymbol{y}_i$, we consider the distribution of

$$T_{(i)}^{c*} = (\widehat{\sigma}_{(i)}^{c*})^{-1}\big\{h(h^{-1}(\xi_i^{c*}, \widehat{\lambda}), \widehat{\lambda}_{(i)}^*) - \hat{\xi}_{(i)}^{EBc*}\big\}, \tag{22}$$

which is denoted by $\mathcal{L}_m^{c*} = \mathcal{L}_m^{c*}(\cdot|\boldsymbol{y}_i)$. Similarly to the unconditional case, we can obtain a conditional prediction interval via the parametric bootstrap approximation.

Theorem 2. *Under Assumptions 1, 2 and 3, we have*

$$\sup_{q\in\mathbb{R}} |\mathcal{L}_m^c(q|\boldsymbol{y}_i) - \mathcal{L}_m^{c*}(q|\boldsymbol{y}_i)| = O_p(m^{-3/2}). \tag{23}$$

The proof of Theorem 2 is given in the Appendix. From Theorem 2, we obtain a conditional prediction interval with second-order accuracy.

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

$$\mathcal{L}_m^{c*}(q_2^c|\boldsymbol{y}_i) - \mathcal{L}_m^{c*}(q_1^c|\boldsymbol{y}_i) = 1 - \alpha,$$

where $\mathcal{L}_m^{c*}(\cdot|\boldsymbol{y}_i)$ is the distribution function of $T_{(i)}^{c*}$. Then, one gets the prediction interval of $h^{-1}(\xi_i, \lambda)$ given by

$$I_m^c = \big[h^{-1}(\hat{\xi}_i^{EB} + q_1^c\widehat{\sigma}_i, \widehat{\lambda}),\ h^{-1}(\hat{\xi}_i^{EB} + q_2^c\widehat{\sigma}_i, \widehat{\lambda})\big]. \tag{24}$$

Under Assumptions 1, 2 and 3, it holds that

$$P\big(h^{-1}(\xi_i, \lambda) \in I_m^c|\boldsymbol{y}_i\big) = 1 - \alpha + O_p(m^{-3/2}). \tag{25}$$

4 Simulation Studies

In this section, we investigate finite performances of the unconditional and conditional prediction intervals suggested in the previous section for the DP and DPL transformations. The performances are examined by Monte Carlo simulation in the case of $\boldsymbol{x}_{ij}'\boldsymbol{\beta} = \mu$ without covariates as treated in Chatterjee, *et al.* (2008).

In the simulation experiments, 1,000 observations for y_{ij} are generated as $y_{ij} = h^{-1}(v_i + \varepsilon_{ij}, \lambda)$, $i = 1, \ldots, m$, $j = 1, \ldots, n_i$, for $m = 10$, $n_i = 5$ and $\lambda = 0$, 0.5 and 1.0, where v_i's and ε_{ij}'s are mutually independently generated from $\mathcal{N}(0,1)$ with $\mu = 0$, $\sigma_v^2 = 1$ and $\sigma_e^2 = 1$. The frequency of the prediction interval 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, we investigate the performances of the unconditional prediction interval and compare it with the naive prediction interval given by

$$\big[h^{-1}(\hat{\xi}_i^{EB} - z_{\alpha/2}\widehat{\sigma}_i, \widehat{\lambda}), h^{-1}(\hat{\xi}_i^{EB} + z_{\alpha/2}\widehat{\sigma}_i, \widehat{\lambda})\big], \tag{26}$$

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. The maximum likelihood estimators are used for the variance components σ_v^2 and σ_e^2. Table 1 reports the coverage probability (CP) and the expected length (EL) of the two unconditional prediction intervals (19) and (26) based on the bootstrap method (BT) and the naive method (NV). From Table 1, it is observed that the naive prediction interval is not appropriate since it does not satisfy the nominal confidence coefficient $1-\alpha=0.95$, while it gives a shorter length than BT. The prediction interval (19) based on BT has the coverage probability close to the nominal level 0.95. This shows that the correction by the bootstrap method works well.

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

	λ	0		0.5		1	
		NV	BT	NV	BT	NV	BT
DPT	CP	91.7	96.1	90.6	94.2	90.6	96.3
	EL	2.64	3.91	2.00	2.40	1.50	5.50
DPLT	CP	91.0	95.4	91.2	95.6	90.7	95.4
	EL	0.32	0.41	0.30	0.45	0.27	0.37

We next investigate a performance of the conditional prediction interval given in (24). The same simulation setup as used above is treated for $\lambda=0.5$ except for fixing initial values of $\boldsymbol{y}_i$'s. We first generate initial observations of $\boldsymbol{y}_i$'s from the model described above for $i=1,\ldots,10$ and fix them. Then, the conditional prediction intervals given $\boldsymbol{y}_i$ are constructed based on the quantiles of the parametric bootstrap samples. The coverage probability (CP) and the expected length (EL) of the conditional prediction interval are reported in Table 2 for TNERM with DPT and DPLT, where the values of $\overline{y}_i$ are the averages $\overline{y}_i$ of the given values $\boldsymbol{y}_i$ for 10 areas. From Table 2, it is revealed that the coverage probabilities of the conditional prediction intervals are close to the nominal level 0.95 for DPT and DPLT. It is interesting to point out that the expected length of the conditional prediction interval for DPT is larger as the value of $\overline{y}_i$ is larger, while the expected length in the case of DPLT is not affected by the value of $\overline{y}_i$. This property of the conditional prediction interval for DPLT is quite different from the unconditional prediction interval.

5 Application to the crop areas data

We now apply the unconditional and conditional prediction intervals to real data. The data we handle is the crop areas data given by Battese, *et al.* (1988), which have been used repeatedly in the literature. Note that the crop (corn) areas data y_{ij}'s are positive and bounded above from 250 hectares. Thus, we can apply the two kinds of transformed nested error regression

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

	area	1	2	3	4	5	6	7	8	9	10
	$\overline{y}_i$	0.44	0.45	1.74	1.91	2.11	2.25	2.47	3.46	3.84	6.58
DPT	CP	95.1	94.9	97.0	95.4	95.5	96.3	97.8	96.2	95.2	93.4
	EL	0.90	0.84	3.33	3.48	3.79	3.95	4.30	5.33	5.23	6.12
	$\overline{y}_i$	0.26	0.38	0.44	0.47	0.49	0.50	0.52	0.60	0.66	0.68
DPLT	CP	95.2	95.7	94.9	96.3	95.5	97.4	95.1	97.7	96.6	96.2
	EL	0.33	0.41	0.42	0.47	0.41	0.59	0.43	0.49	0.39	0.37

models (TNERM) with the dual power transformation (DPT) and the dual power logistic transformation (DPLT).

We begin by applying TNERM with DPT, described as

$$\frac{y_{ij}^{\lambda}-y_{ij}^{-\lambda}}{2\lambda}=\beta_0+\beta_1 x_{1ij}^{\dagger}+\beta_2 x_{2ij}^{\dagger}+v_i+\varepsilon_{ij}, \qquad i=1,\ldots,12 \tag{27}$$

for $x_{1ij}^{\dagger}=\log(x_{1ij})$ and $x_{21ij}^{\dagger}=\log(x_{2ij})$, where x_{1ij}, x_{2ij}, v_i and ε_{ij} are defined around (1). The quantity of interest is the crop areas in the i-th county given by

$$\eta_i=\left(\lambda\xi_i+\sqrt{\lambda^2\xi_i^2+1}\right)^{1/\lambda}, \qquad i=1,\ldots,12$$

where $\xi_i=\beta_0+\beta_1\sum_{j=1}^{n_i}x_{1ij}^{\dagger}/n_i+\beta_2\sum_{j=1}^{n_i}x_{2ij}^{\dagger}/n_i+v_i$. Then the estimates of the parameters $\boldsymbol{\beta}=(\beta_0,\beta_1,\beta_2)'$, σ_v^2, σ_e^2 and λ via the maximum likelihood method are $\widehat{\boldsymbol{\beta}}=(0.26,0.84,-0.05)$, $\widehat{\sigma}_v=0.05$, $\widehat{\sigma}_e=0.14$ and $\widehat{\lambda}=0.00$. These estimates demonstrate a couple of important features of data.

First, the estimate $\widehat{\beta}_2=-0.05$ suggests that the variable x_{2j} does not affect the survey data y_{ij} of corn. This observation seems natural because x_{2j} denotes the numbers of pixels classified as soybeans from the satellite data.

Secondly, the estimate of λ gives $\widehat{\lambda}=0.0$, which recommends the logarithm transformation in model (27). Thus, in analysis of the crop areas data in model (27), we suggest the log-transformed NERM given by

$$\log(y_{ij})=\beta_0+\beta_1 x_{1ij}+\beta_2 x_{2ij}+v_i+\varepsilon_{ij}. \tag{28}$$

Since the model (28) does not contain any transformation parameters, we can obtain the prediction intervals more easily based on the parametric bootstrap method given in Chatterjee, *et al.* (2008). Then, based on 1,000 bootstrap samples, we can construct unconditional and conditional prediction intervals, which are illustrated in Figure 1. From Figure 1, it is revealed

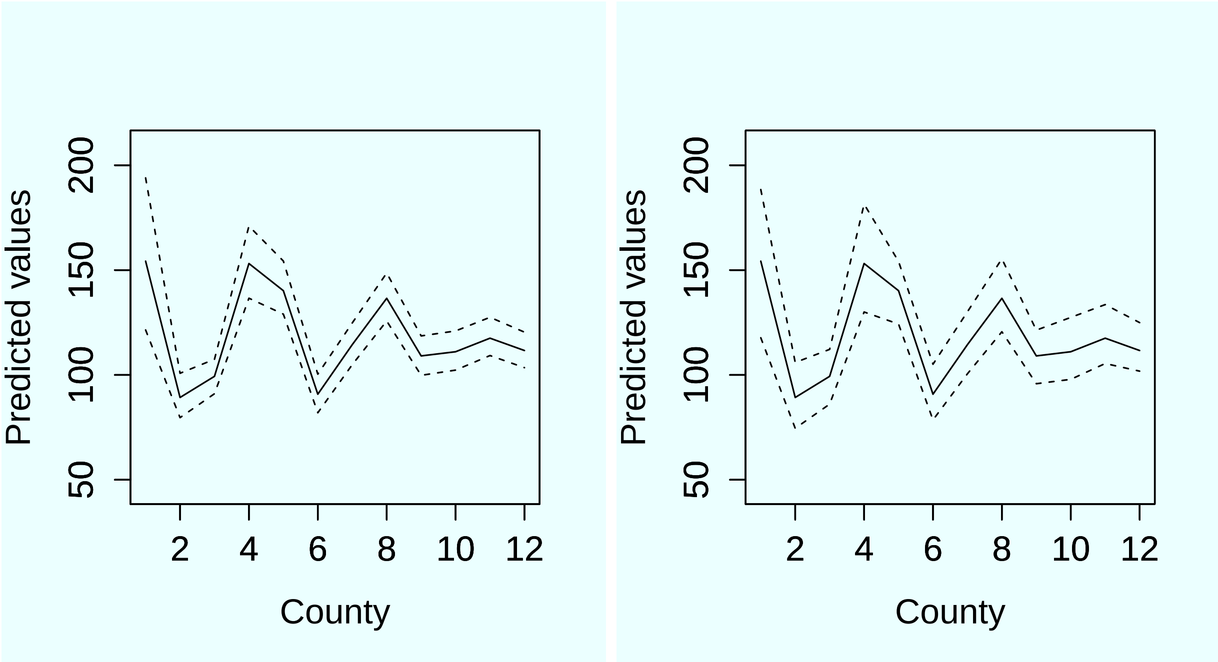

Figure 1: Conditional (right) and Unconditional (left) Prediction Intervals in Model (28) (The solid line denotes TEBLUP and the two dotted lines denote the upper and lower bounds of prediction intervals of each county.)

that the lengths of both prediction intervals are shorter as the sample size is larger. It is also seen that the length of the conditional prediction intervals is larger than that of unconditional ones.

We next try to apply TNERM with the dual power logistic transformation. As explained in Section 1, the crop areas data y_{ij}'s are bounded above from 250 hectares, which means that $100 \times y_{ij}/250$ indicates percentage of crop areas in 250 hectares. Then, the scaled observation $z_{ij} = y_{ij}/250$ lies in the interval $(0, 1)$, and we can apply TNERM with DPLT, described as

$$(2\lambda)^{-1}\left\{\left(\frac{z_{ij}}{1-z_{ij}}\right)^{\lambda}-\left(\frac{z_{ij}}{1-z_{ij}}\right)^{-\lambda}\right\} = \beta_0+\beta_1 \log\left(\frac{x^*_{1ij}}{1-x^*_{1ij}}\right)+\beta_2 \log\left(\frac{x^*_{2ij}}{1-x^*_{2ij}}\right)+v_i+\varepsilon_{ij}, \quad (29)$$

where $x^*_{1ij} = 0.45x_{1ij}/250$ and $x^*_{2ij} = 0.45x_{2ij}/250$. Note that $0.45x_{1ij}$ and $0.45x_{2ij}$ are bounded above from 250. The estimates of parameters by the maximum likelihood method are given by $\widehat{\boldsymbol{\beta}} = (-0.23, 0.85, -0.05)'$, $\widehat{\sigma}_v = 0.11$, $\widehat{\sigma}_e = 0.28$ and $\widehat{\lambda} = 0.37$. Since the estimate of λ is away from 0, this shows that the standard logistic transformation is not appropriate in the framework of model (29). It is noted that the estimate $\hat{\beta}_2$ is close to zero, which implies that the survey data of corn areas are not affected by x_{2ij}, the number of pixels of soybeans. This observation coincides with the analysis based on model (27). Based on 1,000 bootstrap samples, we get the unconditional and conditional prediction intervals, which are illustrated in Figure 2. Similarly to Figure 1, Figure 2 shows that the length of both prediction intervals is shorter as the sample size is larger, and that the conditional prediction intervals are longer than that of the unconditional ones. Comparing Figures 1 and 2, we can see that the lengths of the prediction intervals in model (29) are longer than larger than those in model (28). This may be caused from the prediction intervals based on model (29) involve error in estimation of λ.

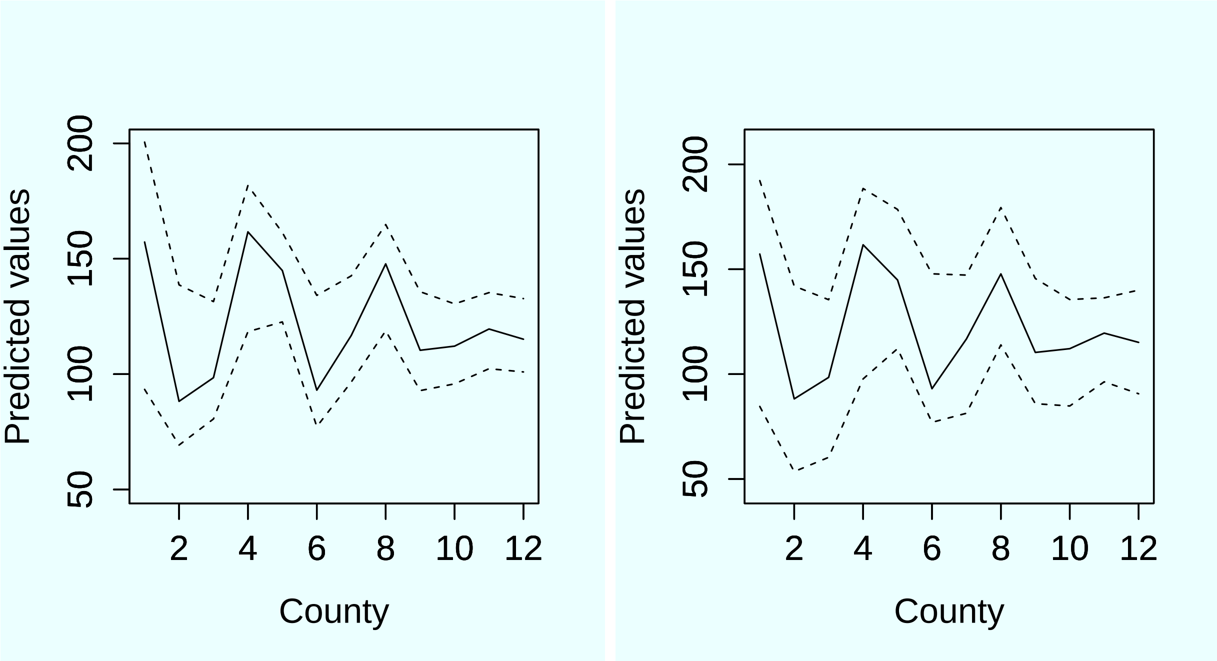

Figure 2: Conditional (right) and Unconditional (left) Prediction Intervals in Model (29) (The solid line denotes TEBLUP and the two dotted lines denote the upper and lower bounds of prediction intervals of each county.)

Finally, we investigate how different the predicted values of $h^{-1}(\xi_i, \lambda)$ are among TEBLUP in (28), TEBLUP in (29) and EBLUP in the non-transformed model given in Battese, *et al.* (1988) defined as

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + v_i + \varepsilon_{ij}. \tag{30}$$

Those predicted values are reported in Table 3 for twelve counties. Although the sample means for small sample sizes seem unreliable, those sample means are much shrunken by EBLUP and the two TEBLUPs. EBLUP and the two TEBLUPs give slightly different predicted values, but perform similarly.

6 Concluding Remarks

In this paper, we have suggested the parametric transformed nested error regression model (TNERM) as a new unit-level model for analysis of positive or bounded data. We have provided the procedures for estimating unknown parameters including the transformation parameter as well as regression coefficients and the variance components. Conditions on the parametric transformation has been derived for consistency of the estimation procedures. The conditions are satisfied by the dual power transformation for positive data and the dual power logistic transformation, which we newly proposed in this paper, for bounded data. The transformed EBLUP (TEBLUP) has been made based on the consistent estimators, and unconditional

Table 3: Predicted Hectares of Corn via EBLUP in NERM and TEBLUP for DPT and DPLT

County	sample size	sample mean	EBLUP in (30)	TEBLUP in (28)	TEBLUP in (29)
Cerro Gordo	1	165.8	155.0	154.3	157.2
Hamilton	1	96.3	89.2	89.2	88.2
Worth	1	76.1	99.1	99.3	98.5
Humboldt	2	150.9	157.7	153.1	161.6
Franklin	3	158.6	144.4	140.2	144.9
Pocahontas	3	102.5	95.3	90.8	93.0
Winnebago	3	112.8	117.3	114.5	116.8
Wright	3	144.3	142.3	136.5	147.8
Webster	4	117.6	111.1	109.1	110.3
Hancock	5	109.4	112.4	111.1	112.1
Kossuth	5	110.3	119.3	117.5	119.6
Hardin	5	114.8	115.5	111.6	115.1

and conditional prediction intervals with second-order accuracy have been constructed based on the parametric bootstrap method. It has been confirmed by simulation that the coverage probability of the suggested prediction intervals is close to the nominal level 0.95. It has been pointed out that the conditional prediction interval given $\boldsymbol{y}_i$ gets wider as the average value $\overline{y}_i$ is larger.

The crop areas data treated in Battese, *et al*. (1988) are positive and bounded above from 250 hectares. We have analyzed the data in the two ways, TNERM with the DPT and DPLT, and we have provided reasonable conditional and unconditional prediction intervals.

Our proposed methodology based on the parametric transformation is regarded as a new framework to cope with small-area data, and we hope further development will be studied from theoretical and practical aspects in statistical inferences.

Appendix

All the lemmas and theorems given in this paper will be proved here. In their proofs, the following fact will be heavily used: Assume that for $i = 1, \ldots, m$, a function $\psi(\boldsymbol{y}_i)$ is independent of $\boldsymbol{y}_j$ for $j \neq i$, and that $\psi(\boldsymbol{y}_i) = O_p(1)$ and $E[\psi(\boldsymbol{y}_i)] = O(1)$. Then, it follows from the Law of Large Numbers (LLN) that

$$\frac{1}{m}\sum_{j=1}^{m} \psi(\boldsymbol{y}_j)\Big|\boldsymbol{y}_i = O_p(1), \quad i = 1, \ldots, m. \tag{31}$$

Moreover, if $E[\psi^2(\boldsymbol{y}_i)] = O(1)$, then from the Central Limit Theorem (CLT), one gets

$$\frac{1}{\sqrt{m}}\Big(\sum_{j=1}^m \psi(\boldsymbol{y}_j) - E\big[\sum_{j=1}^m \psi(\boldsymbol{y}_j)\big|\boldsymbol{y}_i\big]\Big)\Big|\boldsymbol{y}_i = O_p(1), \quad i = 1, \dots, m, \tag{32}$$

where $(\cdot)|\boldsymbol{y}_i$ denotes a random variable given $\boldsymbol{y}_i$.

In the proofs, for notational simplicity, we treat the case of $i = m$ without loss of generality.

A.1 Proof of Lemma 1. Recall that $\widehat{\boldsymbol{\beta}}(\lambda) = \widehat{\boldsymbol{\beta}}(\widehat{\rho}(\lambda), \lambda)$. It is noted that

$$\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}(\rho, \lambda) - \boldsymbol{\beta} + \Big(\frac{\partial\widehat{\boldsymbol{\beta}}(\rho, \lambda)}{\rho}\Big)'(\widehat{\rho}(\lambda) - \rho) + O_p((\widehat{\rho}(\lambda) - \rho)^2). \tag{33}$$

Here $\partial\widehat{\boldsymbol{\beta}}(\rho, \lambda)/\partial\rho$ is expressed as

$$\frac{\partial\widehat{\boldsymbol{\beta}}(\rho, \lambda)}{\partial\rho} = \Big(\sum_{i=1}^m \frac{n_i\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i'}{1 + n_i\rho}\Big)^{-1}\Big(\sum_{i=1}^m \frac{n_i^2\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i'}{(1 + n_i\rho)^2}\Big)\big(\widehat{\boldsymbol{\beta}}(\rho, \lambda) - \widehat{\boldsymbol{\beta}}^{\dagger}(\rho, \lambda)\big),$$

where $\widehat{\boldsymbol{\beta}}^{\dagger}(\rho, \lambda) = \big(\sum_{i=1}^m n_i^2\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i'/(1 + n_i\rho)^2\big)^{-1}\sum_{i=1}^m n_i^2\bar{\boldsymbol{x}}_i z_i(\lambda)/(1 + n_i\rho)^2$. Note that $(\widehat{\boldsymbol{\beta}}^{\dagger}(\rho, \lambda) - \widehat{\boldsymbol{\beta}}(\lambda))|\boldsymbol{y}_m = \boldsymbol{O}_p(m^{-1/2})$ since $(\widehat{\boldsymbol{\beta}}^{\dagger}(\rho, \lambda) - \boldsymbol{\beta})|\boldsymbol{y}_m = \boldsymbol{O}_p(m^{-1/2})$ from (32) and Assumption 2. Thus, $(\partial\widehat{\boldsymbol{\beta}}(\rho, \lambda)/\partial\rho)|\boldsymbol{y}_m = O_p(m^{-1/2})$. Also, $\widehat{\rho} - \rho$ can be expanded as

$$\widehat{\rho}(\lambda) - \rho = \frac{1}{\sigma_e^2}(\widehat{\sigma}_e^2(\lambda) - \sigma_e^2) - \frac{\sigma_v^2}{\sigma_e^4}(\widehat{\sigma}_v^2(\lambda) - \sigma_v^2) + O_p(m^{-1}), \tag{34}$$

which implies that $(\widehat{\rho}(\lambda) - \rho)|\boldsymbol{y}_m = O_p(m^{-1/2})$ and $E[\widehat{\rho}(\lambda) - \rho|\boldsymbol{y}_m] = O_p(m^{-1})$ from Assumptions (A.6) and (A.7). Combining these observations and applying (32) to the first term in the r.h.s. of (33), one gets $(\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta})|\boldsymbol{y}_m = \boldsymbol{O}_p(m^{-1/2})$ under Assumptions 2 and 3.

To show $E[\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}|\boldsymbol{y}_m] = \boldsymbol{O}_p(m^{-1/2})$, from (33), it is sufficient to show that $E[\widehat{\boldsymbol{\beta}}(\rho, \lambda) - \boldsymbol{\beta}|\boldsymbol{y}_m] = O_p(m^{-1})$. Note that $\widehat{\boldsymbol{\beta}}(\rho, \lambda) - \boldsymbol{\beta}$ is rewritten as

$$\begin{aligned}\widehat{\boldsymbol{\beta}}(\rho, \lambda) - \boldsymbol{\beta} &= \Big(\sum_{i=1}^m \frac{n_i\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i'}{1 + n_i\rho}\Big)^{-1}\sum_{i=1}^m \frac{n_i\bar{\boldsymbol{x}}_i\big(z_i(\lambda) - \bar{\boldsymbol{x}}_i'\boldsymbol{\beta}\big)}{1 + n_i\rho} \\ &= \Big(\sum_{i=1}^m \frac{n_i\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i'}{1 + n_i\rho}\Big)^{-1}\Big(\sum_{i=1}^{m-1} \frac{n_i\bar{\boldsymbol{x}}_i\big(z_i(\lambda) - \bar{\boldsymbol{x}}_i'\boldsymbol{\beta}\big)}{1 + n_i\rho} + \frac{n_m\bar{\boldsymbol{x}}_m\big(z_m(\lambda) - \bar{\boldsymbol{x}}_m'\boldsymbol{\beta}\big)}{1 + n_m\rho}\Big).\end{aligned}$$

Noting that $z_1(\lambda), \dots, z_{m-1}(\lambda)$ are independent of $\boldsymbol{y}_m$, $z_i(\lambda) = O_p(1)$ and $\sum_{i=1}^m n_i\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i'/(1 + n_i\rho) = \boldsymbol{O}(m)$ under Assumption 2, one gets

$$E[\widehat{\boldsymbol{\beta}}(\rho, \lambda) - \boldsymbol{\beta}|\boldsymbol{y}_m] = \Big(\sum_{i=1}^m \frac{n_i\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i'}{1 + n_i\rho}\Big)^{-1}\frac{n_m\bar{\boldsymbol{x}}_m\big(z_m(\lambda) - \bar{\boldsymbol{x}}_m'\boldsymbol{\beta}\big)}{1 + n_m\rho} = \boldsymbol{O}_p(m^{-1}).$$

To show the third part, by straightforward calculation, one gets

$$\frac{\partial\widehat{\boldsymbol{\beta}}(\lambda)}{\partial\lambda}=\Big(\sum_{i=1}^m\frac{n_i\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i'}{1+n_i\widehat{\rho}(\lambda)}\Big)^{-1}\sum_{i=1}^m\frac{n_i\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i'}{(1+n_i\widehat{\rho}(\lambda))^2}\big\{\widehat{\boldsymbol{\beta}}(\lambda)-\widehat{\boldsymbol{\beta}}^{\dagger}(\lambda)\big\}\Big(\frac{\partial\widehat{\rho}(\lambda)}{\partial\lambda}\Big)$$
$$+\Big(\sum_{i=1}^m\frac{n_i\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i'}{1+n_i\widehat{\rho}(\lambda)}\Big)^{-1}\sum_{i=1}^m\frac{n_i\bar{\boldsymbol{x}}_iz_{i,\lambda}(\lambda)}{1+n_i\widehat{\rho}(\lambda)},$$

where $\widehat{\boldsymbol{\beta}}^{\dagger}(\lambda)=\big(\sum_{i=1}^m n_i^2\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i'/(1+n_i\widehat{\rho}(\lambda))^2\big)^{-1}\sum_{i=1}^m n_i^2\bar{\boldsymbol{x}}_iz_i(\lambda)/(1+n_i\widehat{\rho}(\lambda))^2$,

$$\frac{\partial\widehat{\rho}(\lambda)}{\partial\lambda}=\frac{1}{\widehat{\sigma}_e^2(\lambda)}\Big(\frac{\partial}{\partial\lambda}\widehat{\sigma}_v^2(\lambda)-\widehat{\rho}(\lambda)\frac{\partial}{\partial\lambda}\widehat{\sigma}_e^2(\lambda)\Big)\quad\text{and}\quad z_{i,\lambda}=\frac{\partial}{\partial\lambda}z_i(\lambda),\quad i=1,\ldots,m.$$

Since $\widehat{\boldsymbol{\beta}}^{\dagger}(\lambda)-\boldsymbol{\beta}=\boldsymbol{O}_p(m^{-1/2})$, we have $\widehat{\boldsymbol{\beta}}(\lambda)-\widehat{\boldsymbol{\beta}}^{\dagger}(\lambda)=\boldsymbol{O}_p(m^{-1/2})$. Also note that that $\partial\widehat{\rho}(\lambda)/\partial\lambda\big|\boldsymbol{y}_m=O_p(1)$ from Assumption (A.8). Then, we have

$$E\Big[\frac{\partial\widehat{\boldsymbol{\beta}}(\lambda)}{\partial\lambda}\Big|\boldsymbol{y}_m\Big]=\Big(\sum_{i=1}^m\frac{n_i\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i'}{1+n_i\rho}\Big)^{-1}\Big(\sum_{i=1}^{m-1}\frac{n_i\bar{\boldsymbol{x}}_iE[z_i(\lambda)]}{1+n_i\rho}+\frac{n_m\bar{\boldsymbol{x}}_mz_m(\lambda)}{1+n_m\rho}\Big)+\boldsymbol{O}_p(m^{-1/2}).$$

Hence, one gets

$$\sqrt{m}\Big(\frac{\partial\widehat{\boldsymbol{\beta}}(\lambda)}{\partial\lambda}-E\Big[\frac{\partial\widehat{\boldsymbol{\beta}}(\lambda)}{\partial\lambda}\Big|\boldsymbol{y}_m\Big]\Big)\Big|\boldsymbol{y}_m$$
$$=\Big(\frac{1}{m}\sum_{i=1}^m\frac{n_i\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i'}{1+n_i\rho}\Big)^{-1}\frac{\sqrt{m-1}}{\sqrt{m}}\frac{1}{\sqrt{m-1}}\sum_{i=1}^{m-1}\frac{n_i\bar{\boldsymbol{x}}_i}{1+n_i\rho}\big\{z_i(\lambda)-E[z_i(\lambda)]\big\}+\boldsymbol{O}_p(1),$$

which is of order $\boldsymbol{O}_p(1)$, since from CLT and Assumption 2,

$$\frac{1}{\sqrt{m-1}}\sum_{i=1}^{m-1}\frac{n_i\bar{\boldsymbol{x}}_i}{1+n_i\rho}\big\{z_i(\lambda)-E[z_i(\lambda)]\big\}=O_p(1).$$

Therefore, Lemma 1 is proved. □

A.2 Proof of Lemma 2. We can easily verify that Assumptions (A.6) and (A.7) are satisfied for the three estimators of $\boldsymbol{\sigma}^2$ based on their stochastic expansions given in Prasad and Rao (1990), Datta and Lahiri (2000) and Das, Jiang and Rao (2004). Thus, we shall check Assumptions (A.8) and (A.9) for $i=m$.

[1] PR estimator. Recall that $\widehat{\boldsymbol{\sigma}}^2_{PR}$ is given in (9). For S_1,

$$\frac{1}{m}\Big(\frac{\partial}{\partial\lambda}S_1(\lambda)\Big)=\frac{2}{m}\boldsymbol{h}(\boldsymbol{Y},\lambda)'(\boldsymbol{I}_N-\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\Big(\frac{\partial}{\partial\lambda}\boldsymbol{h}(\boldsymbol{Y},\lambda)\Big)$$
$$=\frac{2}{m}\sum_{i=1}^m\boldsymbol{h}(\boldsymbol{y}_i,\lambda)'\boldsymbol{M}_i\Big(\frac{\partial}{\partial\lambda}\boldsymbol{h}(\boldsymbol{y}_i,\lambda)\Big)+o_p(1)$$
$$=\frac{2}{m}\sum_{i=1}^{m-1}\boldsymbol{h}(\boldsymbol{y}_i,\lambda)'\boldsymbol{M}_i\Big(\frac{\partial}{\partial\lambda}\boldsymbol{h}(\boldsymbol{y}_i,\lambda)\Big)+\frac{2}{m}\boldsymbol{h}(\boldsymbol{y}_m,\lambda)'\boldsymbol{M}_m\Big(\frac{\partial}{\partial\lambda}\boldsymbol{h}(\boldsymbol{y}_m,\lambda)\Big)+o_p(1),$$

where $\boldsymbol{M}_i = \boldsymbol{I}_{n_i} - \boldsymbol{X}_i(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_i'$. Then from (31) and (32), it follows that

$$\frac{1}{m}E\Big[\frac{\partial}{\partial\lambda}S_1(\lambda)\Big|\boldsymbol{y}_m\Big] = O_p(1) \text{ and } \frac{1}{m}\Big(\frac{\partial}{\partial\lambda}S_1(\lambda) - \frac{1}{m}E\Big[\frac{\partial}{\partial\lambda}S_i(\lambda)\Big|\boldsymbol{y}_m\Big]\Big)\Big|\boldsymbol{y}_m = O_p(m^{-1/2}).$$

For S_2, we can show similar properties since

$$\frac{\partial}{\partial\lambda}S_2(\lambda) = 2\boldsymbol{h}(\boldsymbol{Y},\lambda)'(\boldsymbol{E} - \boldsymbol{E}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{E}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{E})\Big(\frac{\partial}{\partial\lambda}\boldsymbol{h}(\boldsymbol{Y},\lambda)\Big).$$

Thus, Assumptions (A.8) and (A.9) are satisfied.

[2] ML. The ML estimator $\widehat{\boldsymbol{\sigma}}^2_{ML}$ is given in (10). From the implicit function theorem,

$$\frac{\partial}{\partial\lambda}\widehat{\boldsymbol{\sigma}}^2_{ML}(\lambda) = \boldsymbol{I}(\lambda)^{-1}\boldsymbol{J}(\lambda),$$

where

$$\begin{aligned}\boldsymbol{I}(\lambda) =&\begin{pmatrix} I_{11}(\lambda) & I_{12}(\lambda) \\ I_{21}(\lambda) & I_{22}(\lambda)\end{pmatrix} = \begin{pmatrix} \partial L_1(\boldsymbol{\sigma}^2,\lambda)/\partial\sigma_e^2 & \partial L_1(\boldsymbol{\sigma}^2,\lambda)/\partial\sigma_v^2 \\ \partial L_2(\boldsymbol{\sigma}^2,\lambda)/\partial\sigma_e^2 & \partial L_2(\boldsymbol{\sigma}^2,\lambda)/\partial\sigma_v^2\end{pmatrix}\Big|_{\boldsymbol{\sigma}^2=\widehat{\boldsymbol{\sigma}}^2_{ML}(\lambda)}, \\ \boldsymbol{J}(\lambda) =&(J_1(\lambda), J_2(\lambda))' = \Big(\frac{\partial}{\partial\lambda}L_1(\boldsymbol{\sigma}^2,\lambda)\Big|_{\boldsymbol{\sigma}^2=\widehat{\boldsymbol{\sigma}}^2_{ML}(\lambda)}, \frac{\partial}{\partial\lambda}L_2(\boldsymbol{\sigma}^2,\lambda)\Big|_{\boldsymbol{\sigma}^2=\widehat{\boldsymbol{\sigma}}^2_{ML}(\lambda)}\Big)'. \end{aligned} \tag{35}$$

By straightforward calculation, it is shown that

$$\begin{aligned}\frac{\partial}{\partial\lambda}L_1(\boldsymbol{\sigma}^2,\lambda) = \frac{1}{\sigma_e^4}\sum_{i=1}^m\Big(&\boldsymbol{h}(\boldsymbol{y}_i,\lambda) - \boldsymbol{X}_i\widehat{\boldsymbol{\beta}}(\rho,\lambda) - \frac{n_i\rho}{1+n_i\rho}(z_i(\lambda) - \bar{\boldsymbol{x}}_i'\widehat{\boldsymbol{\beta}}(\rho,\lambda))\boldsymbol{j}_i\Big)' \\ &\cdot\Big(\boldsymbol{X}_i + \frac{n_i\rho}{1+n_i\rho}\boldsymbol{j}_i\bar{\boldsymbol{x}}_i'\Big)\Big(\frac{\partial}{\partial\lambda}\widehat{\boldsymbol{\beta}}(\rho,\lambda)\Big),\end{aligned}$$

where

$$\frac{\partial}{\partial\lambda}\widehat{\boldsymbol{\beta}}(\rho,\lambda) = \Big(\sum_{i=1}^m\frac{n_i\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i'}{1+n_i\rho}\Big)^{-1}\sum_{i=1}^m\frac{n_i\bar{\boldsymbol{x}}_i z_{i,\lambda}(\lambda)}{1+n_i\rho},$$

which is of order $\boldsymbol{O}_p(1)$ under Assumption 1 and 2. Then from the above expression, it easily follows that

$$\begin{aligned}\frac{1}{m}J_1(\lambda) =& \sigma_e^{-4}\frac{1}{m}\sum_{i=1}^m\Big(\boldsymbol{h}(\boldsymbol{y}_i,\lambda) - \boldsymbol{X}_i\boldsymbol{\beta} - \frac{n_i\rho}{1+n_i\rho}(z_i(\lambda) - \bar{\boldsymbol{x}}_i'\boldsymbol{\beta})\boldsymbol{j}_i\Big)' \\ &\cdot\Big(\boldsymbol{X}_i + \frac{n_i\rho}{1+n_i\rho}\boldsymbol{j}_i\bar{\boldsymbol{x}}_i'\Big)\Big(\frac{\partial}{\partial\lambda}\widehat{\boldsymbol{\beta}}(\rho,\lambda)\Big) + O_p(m^{-1/2}) \\ =& \Big(\frac{1}{m}\sum_{i=1}^m\boldsymbol{J}_{1i}(\lambda)\Big)'\Big(\frac{\partial}{\partial\lambda}\widehat{\boldsymbol{\beta}}(\rho,\lambda)\Big) + O_p(m^{-1/2}), \quad \text{(say)}.\end{aligned}$$

Since $\boldsymbol{J}_{1i}(\lambda)$, $i = 1,\dots,m$, are mutually independent random vectors $E[\boldsymbol{J}_{1i}(\lambda)] = \boldsymbol{0}$, it is seen that $m^{-1}\sum_{i=1}^m\boldsymbol{J}_{1i}(\lambda)|\boldsymbol{y}_m = O_p(m^{-1/2})$ from (32). Thus, $m^{-1}J_1(\lambda)|\boldsymbol{y}_m = O_p(m^{-1/2})$, namely,

$$m^{-1/2}J_1(\lambda)|\boldsymbol{y}_m = O_p(1).$$

Also, we obtain

$$\frac{\partial}{\partial\lambda}L_2(\boldsymbol{\sigma}^2,\lambda)=\sum_{i=1}^m\frac{2n_i^2}{(\sigma_e^2+n_i\sigma_v^2)^2}\Big(z_i(\lambda)-\bar{\boldsymbol{x}}_i'\widehat{\boldsymbol{\beta}}(\rho,\lambda)\Big)\Big(z_{i,\lambda}(\lambda)-\bar{\boldsymbol{x}}_i'\Big(\frac{\partial}{\partial\lambda}\widehat{\boldsymbol{\beta}}(\rho,\lambda)\Big)\Big),$$

which implies that

$$\begin{aligned}\frac{1}{m}J_2(\lambda)&=\frac{1}{m}\sum_{i=1}^m\frac{2n_i^2}{(\sigma_e^2+n_i\sigma_v^2)^2}\big(z_i(\lambda)-\bar{\boldsymbol{x}}_i'\boldsymbol{\beta}\big)z_{i,\lambda}(\lambda)+O_p(m^{-1/2})\\&=\frac{1}{m}\sum_{i=1}^m J_{2i}(\lambda)+O_p(m^{-1/2}),\quad\text{(say)}.\end{aligned}$$

Since $J_{2i}(\lambda)=O_p(1)$ and it depends only on $\boldsymbol{y}_i$ of $\boldsymbol{Y}$, from (31) and (32), one gets

$$m^{-1}J_2(\lambda)|\boldsymbol{y}_m=O_p(1)\qquad\text{and}\qquad m^{-1/2}\big(J_2(\lambda)-E\big[J_2(\lambda)\big|\boldsymbol{y}_m\big]\big)\big|\boldsymbol{y}_m=O_p(m^{-1/2}).$$

We next evaluate $\boldsymbol{I}(\lambda)$. We here give a proof for $I_{21}(\lambda)$, and we omit proofs for the other elements since they can be similarly proved. By a straightforward calculation,

$$\begin{aligned}\frac{\partial L_2(\boldsymbol{\sigma}^2,\lambda)}{\partial\sigma_v^2}=&-\sum_{i=1}^m\frac{2n_i^2}{(\sigma_e^2+n_i\sigma_v^2)^3}\big(z_i(\lambda)-\bar{\boldsymbol{x}}_i\widehat{\boldsymbol{\beta}}(\rho,\lambda)\big)^2+\sum_{i=1}^m\frac{2n_i^2}{(\sigma_e^2+n_i\sigma_v^2)^3}\\&-\sum_{i=1}^m\frac{2n_i^2}{(\sigma_e^2+n_i\sigma_v^2)^2}\big(z_i(\lambda)-\bar{\boldsymbol{x}}_i\widehat{\boldsymbol{\beta}}(\rho,\lambda)\big)\bar{\boldsymbol{x}}_i'\Big(\frac{\partial\widehat{\boldsymbol{\beta}}(\rho,\lambda)}{\partial\sigma_e^2}\Big),\end{aligned}$$

where

$$\frac{\partial\widehat{\boldsymbol{\beta}}(\rho,\lambda)}{\partial\sigma_e^2}=-\frac{\sigma_v^2}{\sigma_e^4}\Big(\sum_{i=1}^m\frac{n_i\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i'}{1+n_i\rho}\Big)^{-1}\sum_{i=1}^m\frac{n_i\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i'}{(1+n_i\rho)^2}\big(\widehat{\boldsymbol{\beta}}(\rho,\lambda)-\widehat{\boldsymbol{\beta}}^\dagger(\rho,\lambda)\big),$$

which is $\boldsymbol{O}_p(m^{-1/2})$ since $\widehat{\boldsymbol{\beta}}(\rho,\lambda)-\widehat{\boldsymbol{\beta}}^\dagger(\rho,\lambda)=\boldsymbol{O}_p(m^{-1/2})$ as in the proof of Lemma 1. Then,

$$\frac{1}{m}I_{21}(\lambda)=-\frac{1}{m}\sum_{i=1}^m\frac{2n_i^2}{(\sigma_e^2+n_i\sigma_v^2)^3}\big(z_i(\lambda)-\bar{\boldsymbol{x}}_i\boldsymbol{\beta}\big)^2+\frac{1}{m}\sum_{i=1}^m\frac{2n_i^2}{(\sigma_e^2+n_i\sigma_v^2)^3}+O_p(m^{-1/2}),$$

so that, we have

$$\begin{aligned}\frac{1}{m}I_{21}(\boldsymbol{\sigma}^2,\lambda)\Big|\boldsymbol{y}_m=&-\frac{1}{m}\sum_{i=1}^{m-1}\frac{2n_i^2}{(\sigma_e^2+n_i\sigma_v^2)^2}+\frac{1}{m}\sum_{i=1}^m\frac{2n_i^2}{(\sigma_e^2+n_i\sigma_v^2)^3}\\&-\frac{1}{m}\frac{2n_i^2}{(\sigma_e^2+n_m\sigma_v^2)^3}\big(z_m(\lambda)-\bar{\boldsymbol{x}}_m\boldsymbol{\beta}\big)^2+O_p(m^{-1/2})\\=&-\frac{1}{m}\sum_{i=1}^{m-1}\frac{2n_i^2}{(\sigma_e^2+n_i\sigma_v^2)^2}+\frac{1}{m}\sum_{i=1}^m\frac{2n_i^2}{(\sigma_e^2+n_i\sigma_v^2)^3}+O_p(m^{-1/2}),\end{aligned}$$

since $E\big[\big(z_i(\lambda)-\bar{\boldsymbol{x}}_i\boldsymbol{\beta}\big)^2\big]=\sigma_e^2+n_i\sigma_v^2$. This demonstrates that the leading term is of order $O(1)$. Since the other elements of $\boldsymbol{I}(\lambda)$ can be evaluated similarly, we have

$$m^{-1}\boldsymbol{I}(\lambda)|\boldsymbol{y}_i=\boldsymbol{C}(\boldsymbol{\theta})+\boldsymbol{O}_p(m^{-1/2}),$$

where $\boldsymbol{C}(\boldsymbol{\theta})$ is a non-stochastic matrix with bounded entries, i.e. $\boldsymbol{C}(\boldsymbol{\theta}) = \boldsymbol{O}(1)$.Therefore, one gets

$$\frac{\partial}{\partial\lambda}\widehat{\boldsymbol{\sigma}}^2_{ML}(\lambda)\big|\boldsymbol{y}_m = (m^{-1}\boldsymbol{I}(\lambda))^{-1}m^{-1}\boldsymbol{J}(\lambda)\big|\boldsymbol{y}_m = \boldsymbol{O}_p(1),$$

which shows that the ML estimator satisfies Assumption (A.8). Moreover,

$$\begin{aligned}&\sqrt{m}\Big(\frac{\partial}{\partial\lambda}\widehat{\boldsymbol{\sigma}}^2_{ML}(\lambda) - E\Big[\frac{\partial}{\partial\lambda}\widehat{\boldsymbol{\sigma}}^2_{ML}(\lambda)\Big|\boldsymbol{y}_m\Big]\Big)\Big|\boldsymbol{y}_m \\ &\quad = \boldsymbol{C}(\boldsymbol{\theta})^{-1}m^{-1/2}\big(\boldsymbol{J}(\lambda) - E\big[\boldsymbol{J}(\lambda)\big|\boldsymbol{y}_m\big]\big)|\boldsymbol{y}_m + \boldsymbol{O}_p(1),\end{aligned}$$

which is of order $\boldsymbol{O}_p(1)$. Thus, (A.9) is satisfied.

[3] REML Recall that REML is given in (11). From the implicit function theorem,

$$\frac{\partial}{\partial\lambda}\widehat{\boldsymbol{\sigma}}^2_{RML}(\lambda) = \boldsymbol{I}^R(\lambda)^{-1}\boldsymbol{J}(\lambda),$$

where $\boldsymbol{J}(\lambda)$ is defined in (35) and

$$\boldsymbol{I}^R(\lambda) = \begin{pmatrix}\boldsymbol{I}^R_{11}(\lambda) & \boldsymbol{I}^R_{12}(\lambda)\\ \boldsymbol{I}^R_{21}(\lambda) & \boldsymbol{I}^R_{22}(\lambda)\end{pmatrix} = \boldsymbol{I}(\lambda) + \begin{pmatrix}\partial P_1(\boldsymbol{\sigma}^2)/\partial\sigma_e^2 & \partial P_1(\boldsymbol{\sigma}^2)/\partial\sigma_v^2\\ \partial P_2(\boldsymbol{\sigma}^2)/\partial\sigma_e^2 & \partial P_2(\boldsymbol{\sigma}^2)/\partial\sigma_v^2\end{pmatrix}\Big|_{\boldsymbol{\sigma}^2=\widehat{\boldsymbol{\sigma}}^2_{RML}(\lambda)}.$$

for $P_1(\boldsymbol{\sigma}^2) = \mathrm{tr}\,[(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-2}\boldsymbol{X}]$ and $P_2(\boldsymbol{\sigma}^2) = \mathrm{tr}\,[(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Z}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}]$. Then, the result follows if $m^{-1}P_1(\boldsymbol{\sigma}^2)|\boldsymbol{y}_i = O_p(1)$ and $m^{-1}P_2(\boldsymbol{\sigma}^2)|\boldsymbol{y}_i = O_p(1)$, which can be seen from Assumptions 2 and (A.6). □

A.3 Proof of Lemma 3. We begin by demonstrating the consistency of $\widehat{\lambda}$. According the Cramer method explained in Jiang (2010), we show that the equation $F(\widehat{\boldsymbol{\beta}}(\lambda), \widehat{\sigma}_v^2(\lambda), \widehat{\sigma}_e^2(\lambda), \lambda) = 0$ includes a solution which converges to λ in probability. Let

$$g_m(\lambda') = m^{-1}F(\widehat{\boldsymbol{\beta}}(\lambda'), \widehat{\sigma}_v^2(\lambda'), \widehat{\sigma}_e^2(\lambda'), \lambda'),$$

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

$$g(\lambda') = \lim_{m\to\infty} m^{-1}E_\lambda[F(\widehat{\boldsymbol{\beta}}(\lambda'), \widehat{\sigma}_v^2(\lambda'), \widehat{\sigma}_e^2(\lambda'), \lambda')].$$

When $\lambda' = \lambda$, it is noted that $g(\lambda) = 0$, since $g(\lambda) = \lim_{m\to\infty} m^{-1}E_\lambda[F(\boldsymbol{\beta}, \sigma_v^2, \sigma_e^2, \lambda)] = 0$. 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 $m \to \infty$. 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 $(\widehat{\lambda}-\lambda)|\boldsymbol{y}_m=O_p(m^{-1/2})$ in the case of $i=m$. To this end, we expand the equation (12) around λ to get

$$\sqrt{m}(\widehat{\lambda}-\lambda)=-\frac{m^{-1/2}F(\widehat{\boldsymbol{\beta}}(\lambda),\widehat{\sigma}_v^2(\lambda),\widehat{\sigma}_e^2(\lambda),\lambda)}{m^{-1}\big(\frac{\partial}{\partial\lambda}F(\widehat{\boldsymbol{\beta}}(\lambda),\widehat{\sigma}_v^2(\lambda),\widehat{\sigma}_e^2(\lambda),\lambda)\big|_{\lambda=\lambda^*}\big)}, \tag{36}$$

where λ^* is an intermediate value between λ and $\widehat{\lambda}$. For the numerator in (36), from Lemma 1 and Assumption 3, it is seen that

$$\begin{aligned}\frac{1}{\sqrt{m}}F(\widehat{\boldsymbol{\beta}}(\lambda),\widehat{\sigma}_v^2(\lambda),\widehat{\sigma}_e^2(\lambda),\lambda)\Big|\boldsymbol{y}_m&=\frac{1}{\sqrt{m}}F(\boldsymbol{\beta},\sigma_v^2,\sigma_e^2,\lambda)\Big|\boldsymbol{y}_m+O_p(1)\\&=\frac{1}{\sqrt{m}}\sum_{i=1}^m F_i(\boldsymbol{\theta})\Big|\boldsymbol{y}_m+O_p(1),\end{aligned}$$

where $F_i(\boldsymbol{\theta})=-2\log f(\boldsymbol{y}_i;\boldsymbol{\theta})$ for $\boldsymbol{\theta}=(\boldsymbol{\beta},\sigma_v^2,\sigma_e^2,\lambda)$ and the density function $f(\boldsymbol{y}_i;\boldsymbol{\theta})$ of $\boldsymbol{y}_i$. Since $F_1,\ldots,F_m$ are mutually independently distributed with $E[F_i(\boldsymbol{\theta})]=0$, from (32), it is seen that

$$\frac{1}{\sqrt{m}}F(\widehat{\boldsymbol{\beta}}(\lambda),\widehat{\sigma}_v^2(\lambda),\widehat{\sigma}_e^2(\lambda),\lambda)\Big|\boldsymbol{y}_m=O_p(1).$$

For the denominator in (36), it follows from the consistency of $\widehat{\lambda}$ that

$$m^{-1}\big(\frac{\partial}{\partial\lambda}F(\widehat{\boldsymbol{\beta}}(\lambda),\widehat{\sigma}_v^2(\lambda),\widehat{\sigma}_e^2(\lambda),\lambda)\big|_{\lambda=\lambda^*}\big)=m^{-1}\frac{\partial}{\partial\lambda}F(\widehat{\boldsymbol{\beta}}(\lambda),\widehat{\sigma}_v^2(\lambda),\widehat{\sigma}_e^2(\lambda),\lambda)\big(1+o_p(1)\big).$$

By straightforward calculation, it can be seen from Lemma 1 and Assumption 3 that

$$\begin{aligned}&\frac{1}{m}\Big(\frac{\partial}{\partial\lambda}F(\widehat{\boldsymbol{\beta}}(\lambda),\widehat{\sigma}_v^2(\lambda),\widehat{\sigma}_e^2(\lambda),\lambda)\Big)\\&=\Big\{-\frac{1}{\sigma_e^4}\Big(\frac{\partial}{\partial\lambda}\widehat{\sigma}_e^2(\lambda)\Big)\frac{1}{m}\sum_{i=1}^m(\boldsymbol{h}(\boldsymbol{y}_i,\lambda)-\boldsymbol{X}_i\boldsymbol{\beta})'\boldsymbol{V}_i(\rho)^{-1}\Big(\frac{\partial}{\partial\lambda}\boldsymbol{h}(\boldsymbol{y}_i,\lambda)\Big)\\&+\sigma_e^{-2}\frac{1}{m}\sum_{i=1}^m\Big(\frac{\partial}{\partial\lambda}\boldsymbol{h}(\boldsymbol{y}_i,\lambda)-\boldsymbol{X}_i\Big(\frac{\partial}{\partial\lambda}\widehat{\boldsymbol{\beta}}(\lambda)\Big)\Big)'\boldsymbol{V}_i(\rho)^{-1}\Big(\frac{\partial}{\partial\lambda}\boldsymbol{h}(\boldsymbol{y}_i,\lambda)\Big)\\&+\sigma_e^{-2}\frac{1}{m}\sum_{i=1}^m(\boldsymbol{h}(\boldsymbol{y}_i,\lambda)-\boldsymbol{X}_i\boldsymbol{\beta})'\Big(\frac{\partial}{\partial\lambda}\boldsymbol{V}_i(\widehat{\rho}(\lambda))^{-1}\Big)\Big(\frac{\partial}{\partial\lambda}\boldsymbol{h}(\boldsymbol{y}_i,\lambda)\Big)\\&+\sigma_e^{-2}\frac{1}{m}\sum_{i=1}^m(\boldsymbol{h}(\boldsymbol{y}_i,\lambda)-\boldsymbol{X}_i\boldsymbol{\beta})'\boldsymbol{V}_i(\rho)^{-1}\Big(\frac{\partial^2}{\partial\lambda^2}\boldsymbol{h}(\boldsymbol{y}_i,\lambda)\Big)+\frac{1}{m}\Big(\frac{\partial}{\partial\lambda}J(\boldsymbol{Y},\lambda)\Big)\Big\}\big(1+o_p(1)\big)\\&=(K_1+K_2+K_3+K_4+K_5)(1+o_p(1)).\quad\text{(say)}\end{aligned}$$

We shall evaluate the terms $K_1, \ldots, K_5$ under Assumption 1. It is easy to see that $K_4|\boldsymbol{y}_m = O_p(1)$ and $K_5|\boldsymbol{y}_m = O_p(1)$ by (31). Similarly under Assumptions 1 and 3, we have $K_1 = O_p(1)$ by (31). To evaluate K_2, the expression is rewritten as

$$\begin{aligned} K_2 =&\sigma_e^{-2}\frac{1}{m}\sum_{i=1}^m\Big(\frac{\partial}{\partial\lambda}\boldsymbol{h}(\boldsymbol{y}_i,\lambda)\Big)'\boldsymbol{V}_i(\rho)^{-1}\Big(\frac{\partial}{\partial\lambda}\boldsymbol{h}(\boldsymbol{y}_i,\lambda)\Big) \\ &+\sigma_e^{-2}\Big(\frac{\partial}{\partial\lambda}\widehat{\boldsymbol{\beta}}(\lambda)\Big)'\Big(\frac{1}{m}\sum_{i=1}^m\boldsymbol{X}_i\boldsymbol{V}_i(\rho)^{-1}\Big(\frac{\partial}{\partial\lambda}\boldsymbol{h}(\boldsymbol{y}_i,\lambda)\Big)\Big), \end{aligned}$$

where $(\partial\widehat{\boldsymbol{\beta}}(\lambda)/\partial\lambda)|\boldsymbol{y}_m = \boldsymbol{O}_p(1)$ from Lemma 1. Then from (31), $K_2|\boldsymbol{y}_m = O_p(1)$. For K_3, it is observed that the each element of $(\partial\boldsymbol{V}_i(\widehat{\rho}(\lambda))^{-1}/\partial\lambda)|\boldsymbol{y}_m$ is of order $O_p(1)$, since $(\widehat{\rho}(\lambda)/\partial\lambda)|\boldsymbol{y}_m = O_p(1)$ under Assumption 3. Furthermore, the expression of K_3 reduces to $K_3 = \sigma_e^{-2}\mathrm{tr}\,[\boldsymbol{K}_3^*]$, where

$$\boldsymbol{K}_3^* = \frac{1}{m}\sum_{i=1}^m\Big(\frac{\partial}{\partial\lambda}\boldsymbol{h}(\boldsymbol{y}_i,\lambda)\Big)\big(\boldsymbol{h}(\boldsymbol{y}_i,\lambda)-\boldsymbol{X}_i\boldsymbol{\beta}\big)'\cdot\Big(\frac{\partial}{\partial\lambda}\boldsymbol{V}_i(\widehat{\rho}(\lambda))^{-1}\Big).$$

From (31) and Assumption 1, we have

$$\frac{1}{m}\sum_{i=1}^m\Big(\frac{\partial}{\partial\lambda}\boldsymbol{h}(\boldsymbol{y}_i,\lambda)\Big)\big(\boldsymbol{h}(\boldsymbol{y}_i,\lambda)-\boldsymbol{X}_i\boldsymbol{\beta}\big)'|\boldsymbol{y}_m = \boldsymbol{O}_p(1),$$

so that $\boldsymbol{K}_3^*|\boldsymbol{y}_m = \boldsymbol{O}_p(1)$. Since $\boldsymbol{K}_3^*$ is an $n_i \times n_i$ matrix, it follows that $K_3 = O_p(1)$. These observations show that the denominator in (36) is of order $O_p(1)$. Hence, one gets $\sqrt{m}(\widehat{\lambda} - \lambda)|\boldsymbol{y}_m = O_p(1)$, namely, $(\widehat{\lambda} - \lambda)|\boldsymbol{y}_m = O_p(m^{-1/2})$.

Finally, we show that $E[\widehat{\lambda} - \lambda|\boldsymbol{y}_m] = O_p(m^{-1})$. Evaluating the term in (36) more precisely based on the fact that $(\widehat{\lambda} - \lambda)|\boldsymbol{y}_m = O_p(m^{-1/2})$, we can approximate $\widehat{\lambda} - \lambda$ stochastically as

$$\widehat{\lambda} - \lambda = -\frac{F(\widehat{\boldsymbol{\beta}}(\lambda), \widehat{\sigma}_v^2(\lambda), \widehat{\sigma}_e^2(\lambda), \lambda)}{\frac{\partial}{\partial\lambda}F(\widehat{\boldsymbol{\beta}}(\lambda), \widehat{\sigma}_v^2(\lambda), \widehat{\sigma}_e^2(\lambda), \lambda)} + O_p(m^{-1}).$$

Let $M = E[\partial F(\widehat{\boldsymbol{\beta}}(\lambda), \widehat{\sigma}_v^2(\lambda), \widehat{\sigma}_e^2(\lambda), \lambda)/\partial\lambda]$, which is of order $O(m)$. From Lemma 1 and Assumption 3, it easily follows that

$$\frac{1}{m}\Big(\frac{\partial}{\partial\lambda}F(\widehat{\boldsymbol{\beta}}(\lambda), \widehat{\sigma}_v^2(\lambda), \widehat{\sigma}_e^2(\lambda), \lambda)\Big) = \frac{M}{m} + O_p(m^{-1/2}).$$

Then, one gets $\widehat{\lambda} - \lambda = -M^{-1}F(\widehat{\boldsymbol{\beta}}(\lambda), \widehat{\sigma}_v^2(\lambda), \widehat{\sigma}_e^2(\lambda), \lambda) + O_p(m^{-1})$, so that

$$E[\widehat{\lambda} - \lambda] = -M^{-1}E[F(\widehat{\boldsymbol{\beta}}(\lambda), \widehat{\sigma}_v^2(\lambda), \widehat{\sigma}_e^2(\lambda), \lambda)] + O(m^{-1}).$$

Note that $m^{-1}F(\widehat{\boldsymbol{\beta}}(\lambda), \widehat{\sigma}_v^2(\lambda), \widehat{\sigma}_e^2(\lambda), \lambda)$ is evaluated as

$$
\begin{aligned}
&m^{-1}F(\widehat{\boldsymbol{\beta}}(\lambda), \widehat{\sigma}_v^2(\lambda), \widehat{\sigma}_e^2(\lambda), \lambda) \\
&\quad = m^{-1}F(\boldsymbol{\beta}, \sigma_v^2, \sigma_e^2, \lambda) + \sigma_e^{-2}(\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta})' \frac{1}{m}\sum_{i=1}^m \boldsymbol{V}_i(\rho)^{-1}\Big(\frac{\partial}{\partial\lambda}\boldsymbol{h}(\boldsymbol{y}_i, \lambda)\Big) \\
&\quad + \sigma_e^{-2}\frac{1}{m}\sum_{i=1}^m (\boldsymbol{h}(\boldsymbol{y}_i, \lambda) - \boldsymbol{X}_i\boldsymbol{\beta})'\big\{\boldsymbol{V}_i(\widehat{\rho}(\lambda))^{-1} - \boldsymbol{V}_i(\rho)^{-1}\big\}\Big(\frac{\partial}{\partial\lambda}\boldsymbol{h}(\boldsymbol{y}_i, \lambda)\Big) \\
&\quad + \Big(\frac{1}{\widehat{\sigma}_e^2(\lambda)} - \frac{1}{\sigma_e^2}\Big)\frac{1}{m}\sum_{i=1}^m (\boldsymbol{h}(\boldsymbol{y}_i, \lambda) - \boldsymbol{X}_i\boldsymbol{\beta})'\boldsymbol{V}_i(\rho)^{-1}\Big(\frac{\partial}{\partial\lambda}\boldsymbol{h}(\boldsymbol{y}_i, \lambda)\Big) + O_p(1).
\end{aligned}
$$

From Assumption (A.7), it is easy to see that $E[\widehat{\sigma}_e^{-2}(\lambda) - \sigma_e^{-2}|\boldsymbol{y}_m] = O_p(m^{-1})$ and $E\big[\boldsymbol{V}_i(\widehat{\rho}(\lambda))^{-1} - \boldsymbol{V}_i(\rho)^{-1}\big|\boldsymbol{y}_m\big] = \boldsymbol{O}_p(m^{-1})$, which conclude that $E\big[m^{-1}F(\widehat{\boldsymbol{\beta}}(\lambda), \widehat{\sigma}_v^2(\lambda), \widehat{\sigma}_e^2(\lambda), \lambda)\big|\boldsymbol{y}_m\big] = O_p(m^{-1})$. Therefore, the proof is complete. □

A.4 Proof of Lemma 4. From Lemma 3, we need to establish the results for $\widehat{\boldsymbol{\beta}}(\widehat{\lambda})$ and $\widehat{\boldsymbol{\sigma}}^2(\widehat{\lambda})$. Let $i = m$. From Lemmas 1 and 3, we have

$$
\begin{aligned}
(\widehat{\boldsymbol{\beta}}(\widehat{\lambda}) - \boldsymbol{\beta})|\boldsymbol{y}_m &= \Big(\frac{\partial}{\partial\lambda}\widehat{\boldsymbol{\beta}}(\lambda)\Big)(\widehat{\lambda} - \lambda)\Big|\boldsymbol{y}_m + \boldsymbol{O}_p(m^{-1}) \\
&= \Big(\frac{\partial}{\partial\lambda}\widehat{\boldsymbol{\beta}}(\lambda) - E\Big[\frac{\partial}{\partial\lambda}\widehat{\boldsymbol{\beta}}(\lambda)\Big|\boldsymbol{y}_m\Big]\Big)(\widehat{\lambda} - \lambda)\Big|\boldsymbol{y}_m \\
&\quad + E\Big[\frac{\partial}{\partial\lambda}\widehat{\boldsymbol{\beta}}(\lambda)\Big|\boldsymbol{y}_m\Big](\widehat{\lambda} - \lambda)\Big|\boldsymbol{y}_m + \boldsymbol{O}_p(m^{-1}) \\
&= E\Big[\frac{\partial}{\partial\lambda}\widehat{\boldsymbol{\beta}}(\lambda)\Big|\boldsymbol{y}_m\Big](\widehat{\lambda} - \lambda)\big|\boldsymbol{y}_m + \boldsymbol{O}_p(m^{-1}),
\end{aligned}
$$

since $E[\partial\widehat{\boldsymbol{\beta}}(\lambda)/\partial\lambda|\boldsymbol{y}_m] = \boldsymbol{O}_p(1)$. Then, one gets $(\widehat{\boldsymbol{\beta}}(\widehat{\lambda}) - \boldsymbol{\beta})|\boldsymbol{y}_m = \boldsymbol{O}_p(m^{-1/2})$ and $E[\widehat{\boldsymbol{\beta}}(\widehat{\lambda}) - \boldsymbol{\beta}|\boldsymbol{y}_m] = \boldsymbol{O}_p(m^{-1})$ from Lemmas 1 and 3. Similarly, the results for $\widehat{\boldsymbol{\sigma}}^2(\widehat{\lambda})$ follow from Lemma 3 and Assumption 3, since

$$
(\widehat{\boldsymbol{\sigma}}^2(\widehat{\lambda}) - \boldsymbol{\sigma}^2)\big|\boldsymbol{y}_m = E\Big[\frac{\partial}{\partial\lambda}\widehat{\boldsymbol{\sigma}}^2(\lambda)\Big|\boldsymbol{y}_m\Big](\widehat{\lambda} - \lambda)\big|\boldsymbol{y}_m + \boldsymbol{O}_p(m^{-1}),
$$

where $E[\partial\widehat{\boldsymbol{\sigma}}^2(\lambda)/\partial\lambda\big|\boldsymbol{y}_m] = \boldsymbol{O}_p(1)$. Therefore, the proof is complete. □

A.5 Proof of Theorem 1. It is first noted that in the proof, the capital C, with or without suffix, means a generic constant. If $\mathcal{L}_m(q)$ is expanded as

$$
\mathcal{L}_m(q) = \Phi(q) + m^{-1}\gamma(q, \boldsymbol{\theta}) + O_p(m^{-3/2}), \tag{37}
$$

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

$$\begin{aligned}\mathcal{L}_m^*(q)-\mathcal{L}_m(q)=&m^{-1}\big\{\gamma(q,\widehat{\boldsymbol{\theta}})-\gamma(q,\boldsymbol{\theta})\big\}+O_p(m^{-3/2})\\=&m^{-1}\Big(\frac{\partial\gamma(q,\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\Big)'(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta})+O_p(m^{-3/2}),\end{aligned}\tag{38}$$

which establishes the result given in Theorem 1. Hence, we shall show the expansion (37) through the following steps.

(Step 1) Expansion of $\mathcal{L}_m(q)$. Since the inequality $\widehat{\sigma}_i^{-1}\big\{h(h^{-1}(\xi_i,\lambda),\widehat{\lambda})-\hat{\xi}_i^{EB}\big\}\leq q$ for any $q\in\mathbb{R}$ is equivalently rewritten as $h^{-1}(\xi_i,\lambda)\leq h^{-1}(\hat{\xi}_i^{EB}+q\widehat{\sigma}_i,\widehat{\lambda})$, we have

$$\begin{aligned}\mathcal{L}_m(q)&=P\big[\widehat{\sigma}_i^{-1}\big\{h(h^{-1}(\xi_i,\lambda),\widehat{\lambda})-\hat{\xi}_i^{EB}\big\}\leq q\big]\\&=E\Big(P\big[\sigma_i^{-1}(\xi_i-\hat{\xi}_i(\boldsymbol{\theta}))\leq\sigma_i^{-1}\big\{h(h^{-1}(\hat{\xi}_i^{EB}+q\widehat{\sigma}_i,\widehat{\lambda}),\lambda)-\hat{\xi}_i(\boldsymbol{\theta})\big\}\big]\Big|\boldsymbol{Y}\Big)\\&=E\big[\Phi(q+R(q,\boldsymbol{Y}))\big],\end{aligned}$$

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

$$R(q,\boldsymbol{Y})=\sigma_i^{-1}\big\{h(h^{-1}(\hat{\xi}_i^{EB}+q\widehat{\sigma}_i,\widehat{\lambda}),\lambda)-\hat{\xi}_i(\boldsymbol{\theta})\big\}-q.$$

For the standard normal density function $\phi(\cdot)$, the first and second derivatives are written as $\phi'(x)=-x\phi(x) and\ \phi''(x)=(x^2-1)\phi(x)$ for $x\in\mathbb{R}$. The Taylor expansion is applied to get

$$\begin{aligned}\mathcal{L}_m(q)&=\Phi(q)+\phi(q)E[R(q,\boldsymbol{Y})]-\frac{1}{2}q\phi(q)E[R^2(q,\boldsymbol{Y})]\\&\quad+\frac{1}{2}E\Big[\int_q^{q+R(q,\boldsymbol{Y})}\big(q+R(q,\boldsymbol{Y})-x\big)^2(x^2-1)\phi(x)dx\Big]\\&=\Phi(q)+\phi(q)t_1(q)-\frac{1}{2}q\phi(q)t_2(q)+t_3(q),\end{aligned}$$

where $t_1(q)=E[R]$, $t_2(q)=E[R^2]$ and $t_3(q)=2^{-1}E[\int_q^{q+R}(q+R-x)^2(x^2-1)\phi(x)dx]$ for $R=R(q,\boldsymbol{Y})$. Note that $0\leq|q+R-x|\leq|R|$ and $(x^2-1)\phi(x)\leq2\phi(\sqrt{3})$ for $x\in(q,q+R)$. Then,

$$E\Big[\int_q^{q+R}(q+R-x)^2(x^2-1)\phi(x)dx\Big]\leq E\Big[R^2\int_q^{q+R}2\phi(\sqrt{3})dx\Big]\leq C_1E\big[|R|^3\big],$$

which implies that

$$\mathcal{L}_m(q)=\Phi(q)+\phi(q)t_1(q)-\frac{1}{2}q\phi(q)t_2(q)+O(E\big[|R|^3\big]).\tag{39}$$

(Step 2) Expansion of $R=R(q,\boldsymbol{Y})$. We shall show that $R=R(q,\boldsymbol{Y})=O_p(m^{-1/2})$ based on an expansion of R. It follows from this property that $\sup_{q\in\mathbb{R}}t_3(q)=O(m^{-3/2})$ and $\sup_{q\in\mathbb{R}}t_2(q)=O(m^{-1})$. Let $Q=h^{-1}(\hat{\xi}_i^{EB}+q\widehat{\sigma}_i,\widehat{\lambda})-h^{-1}(\hat{\xi}_i(\boldsymbol{\theta})+q\sigma_i,\lambda)$. Then,

$$\begin{aligned}h(h^{-1}(\hat{\xi}_i^{EB}+q\widehat{\sigma}_i,\widehat{\lambda}),\lambda)=&\hat{\xi}_i(\boldsymbol{\theta})+q\sigma_i+h_x(h^{-1}(\hat{\xi}_i(\boldsymbol{\theta})+q\sigma_i,\lambda),\lambda)Q\\&+\frac{1}{2}h_{xx}(h^{-1}(\hat{\xi}_i(\boldsymbol{\theta})+q\sigma_i,\lambda),\lambda)Q^2\\&+\frac{1}{2}\int_a^{a+Q}(a+Q-x)^2h_{xxx}(x,\lambda)dx,\end{aligned}$$

where $a = h^{-1}(\hat{\xi}_i(\boldsymbol{\theta}) + q\sigma_i, \lambda)$. Since $0 \le |q + Q - x| \le |Q|$ for $x \in (a, a + Q)$, we have

$$\Big| \int_a^{a+Q} (a + Q - x)^2 h_{xxx}(x, \lambda) dx \Big| \le Q^2 \Big| h_{xx}(a + Q, \lambda) - h_{xx}(Q, \lambda) \Big|.$$

It is here noted that $Q = O_p(m^{-1/2})$, which will be shown in (Step 3) below. Then it follows from Assumption 1 that $h_x(a + Q, \lambda)$, $h_x(Q, \lambda)$ and $h_{xx}(a, \lambda)$ are $O_p(1)$. Thus,

$$h(h^{-1}(\hat{\xi}_i^{EB} + q\widehat{\sigma}_i, \widehat{\lambda}), \lambda) = \hat{\xi}_i(\boldsymbol{\theta}) + q\sigma_i + g(\boldsymbol{y}_i, \boldsymbol{\theta})Q + O_p(m^{-1}),$$

for $g(\boldsymbol{y}_i, \boldsymbol{\theta}) = h_x(h^{-1}(\hat{\xi}_i(\boldsymbol{\theta}) + q\sigma_i, \lambda), \lambda)$. Since $g(\boldsymbol{y}_i, \boldsymbol{\theta}) = O_p(1)$, it can be observed that $g(\boldsymbol{y}_i, \boldsymbol{\theta})Q = O_p(m^{-1/2})$, which results in

$$R(q, \boldsymbol{Y}) = \sigma_i^{-1} g(\boldsymbol{y}_i, \boldsymbol{\theta})Q + O_p(m^{-1}). \tag{40}$$

Also, the expectation of $R(q, \boldsymbol{Y})$ is evaluated as

$$E[R(q, \boldsymbol{Y})] = \sigma_i^{-1} E\big[g(\boldsymbol{y}_i, \boldsymbol{\theta}) E(Q|\boldsymbol{y}_i)\big] + O(m^{-1}). \tag{41}$$

(Step 3) Evaluation of Q and $E[Q|\boldsymbol{y}_i]$. To get the expansion (37), it is sufficient to show that $Q = O_p(m^{-1/2})$ and $E[Q|\boldsymbol{y}_i] = O_p(m^{-1})$ from (41). To this end, decompose Q as $Q = Q_1 + Q_2$, where

$$Q_1 = h^{-1}(\hat{\xi}_i^{EB} + q\widehat{\sigma}_i, \lambda) - h^{-1}(\hat{\xi}_i(\boldsymbol{\theta}) + q\sigma_i, \lambda), \tag{42}$$
$$Q_2 = h^{-1}(\hat{\xi}_i^{EB} + q\widehat{\sigma}_i, \widehat{\lambda}) - h^{-1}(\hat{\xi}_i^{EB} + q\widehat{\sigma}_i, \lambda). \tag{43}$$

From (42), Q_1 is expanded as

$$Q_1 = h_x^{-1}(\hat{\xi}_i(\boldsymbol{\theta}) + q\sigma_i, \lambda)U + h_{xx}^{-1}(\hat{\xi}_i(\boldsymbol{\theta}) + q\sigma_i, \lambda)U^2 + \frac{1}{2}\int_b^{b+U} (b + Q - x)^2 h_{xxx}^{-1}(x, \lambda) dx,$$

where $U = \hat{\xi}_i^{EB} - \hat{\xi}_i(\boldsymbol{\theta}) + q(\widehat{\sigma}_i - \sigma_i)$ and $b = \hat{\xi}_i(\boldsymbol{\theta}) + q\sigma_i$. It is here noted that

$$U|\boldsymbol{y}_i = O_p(m^{-1/2}) \quad \text{and} \quad E[U|\boldsymbol{y}_i] = O_p(m^{-1}), \tag{44}$$

which will be shown in (Step 4) below. Then, it follows that the last two terms of the expansion of Q_1 are $O_p(m^{-1})$ given $\boldsymbol{y}_i$, and $Q_1|\boldsymbol{y}_i = O_p(m^{-1/2})$ by the similar argument. Thus,

$$E[Q_1|\boldsymbol{y}_i] = h_x^{-1}(\hat{\xi}_i(\boldsymbol{\theta}) + q\sigma_i, \lambda)E(U|\boldsymbol{y}_i) + O_p(m^{-1}) = O_p(m^{-1}).$$

Also, Q_2 is expanded as

$$Q_2 = h_\lambda^{-1}(\hat{\xi}_i^{EB} + q\widehat{\sigma}_i, \lambda)(\widehat{\lambda} - \lambda) + \frac{1}{2} h_{\lambda\lambda}^{-1}(\hat{\xi}_i^{EB} + q\widehat{\sigma}_i, \lambda^*)(\widehat{\lambda} - \lambda)^2,$$

where λ^* is intermediate value between λ and $\widehat{\lambda}$. It can be observed that $h_\lambda^{-1}(\hat{\xi}_i^{EB}+q\widehat{\sigma}_i,\lambda)|\boldsymbol{y}_i = O_p(1)$, $h_{\lambda\lambda}^{-1}(\hat{\xi}_i^{EB}+q\widehat{\sigma}_i,\lambda^*)|\boldsymbol{y}_i = O_p(1)$ under Assumption 1 and $(\widehat{\lambda}-\lambda)|\boldsymbol{y}_i = O_p(m^{-1/2})$ from Lemma 4. Thus, $Q_2|\boldsymbol{y}_i = O_p(m^{-1/2})$ and

$$\begin{aligned}E[Q_2|\boldsymbol{y}_i] &= E\big[h_\lambda^{-1}(\hat{\xi}_i^{EB}+q\widehat{\sigma}_i,\lambda)(\widehat{\lambda}-\lambda)\big|\boldsymbol{y}_i\big] + O_p(m^{-1})\\ &= E\big[\big\{h_\lambda^{-1}(\hat{\xi}_i^{EB}+q\widehat{\sigma}_i,\lambda) - h_\lambda^{-1}(\hat{\xi}_i(\boldsymbol{\theta})+q\sigma_i,\lambda)\big\}(\widehat{\lambda}-\lambda)\big|\boldsymbol{y}_i\big]\\ &\quad + h_\lambda^{-1}(\hat{\xi}_i(\boldsymbol{\theta})+q\sigma_i,\lambda)E(\widehat{\lambda}-\lambda|\boldsymbol{y}_i) + O_p(m^{-1})\\ &= O_p(m^{-1}),\end{aligned}$$

since $h_\lambda^{-1}(\hat{\xi}_i^{EB}+q\widehat{\sigma}_i,\lambda) - h_\lambda^{-1}(\hat{\xi}_i(\boldsymbol{\theta})+q\sigma_i,\lambda)$ given $\boldsymbol{y}_i$ is $O_p(m^{-1/2})$, which can be verified by $h_{\lambda x}^{-1}(\hat{\xi}_i(\boldsymbol{\theta})+q\sigma_i,\lambda) = O_p(1)$ and Lemma 4.

(Step 4) Evaluation of $U|\boldsymbol{y}_i$ and $E[U|\boldsymbol{y}_i]$. It remains to show that $U|\boldsymbol{y}_i = O_p(m^{-1/2})$ and $E(U|\boldsymbol{y}_i) = O_p(m^{-1})$, for which it is sufficient to show that both $(\hat{\xi}_i^{EB}-\hat{\xi}_i(\boldsymbol{\theta}))|\boldsymbol{y}_i$ and $(\widehat{\sigma}_i-\sigma_i)|\boldsymbol{y}_i$ are $O_p(m^{-1/2})$ and the conditional expectation given $\boldsymbol{y}_i$ is $O_p(m^{-1})$. Recall that $U = \hat{\xi}_i^{EB}-\hat{\xi}_i(\boldsymbol{\theta}) + q(\widehat{\sigma}_i-\sigma_i)$. First, $\hat{\xi}_i^{EB}-\hat{\xi}_i(\boldsymbol{\theta})$ is rewritten as

$$\begin{aligned}\hat{\xi}_i^{EB}-\hat{\xi}_i(\boldsymbol{\theta}) = \bar{\boldsymbol{x}}_i'(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}) + \frac{\widehat{\rho}n_i}{1+\widehat{\rho}n_i}(z_i(\widehat{\lambda})-z_i(\lambda)-\bar{\boldsymbol{x}}_i'(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}))\\ + \Big(\frac{\widehat{\rho}n_i}{1+\widehat{\rho}n_i}-\frac{\rho n_i}{1+\rho n_i}\Big)(z_i(\lambda)-\bar{\boldsymbol{x}}'\boldsymbol{\beta}).\end{aligned}$$

Note that given $\boldsymbol{y}_i$, $z_i(\widehat{\lambda})-z_i(\lambda) = z_{i,\lambda}(\lambda)(\widehat{\lambda}-\lambda)+O_p(m^{-1})$ and

$$\frac{\widehat{\rho}n_i}{1+\widehat{\rho}n_i} = \frac{\rho n_i}{1+\rho n_i} + \frac{n_i}{(1+\rho n_i)^2}(\widehat{\rho}-\rho)+O_p(m^{-1}).$$

Further, from Lemma 4 and a similar expansion as in (34), it follows that $(\widehat{\rho}-\rho)|\boldsymbol{y}_i = O_p(m^{-1/2})$ and $E(\widehat{\rho}-\rho|\boldsymbol{y}_i) = O_p(m^{-1})$. Hence, one gest $(\hat{\xi}_i^{EB}-\hat{\xi}_i(\boldsymbol{\theta}))|\boldsymbol{y}_i = O_p(m^{-1/2})$ and

$$\begin{aligned}&E[\hat{\xi}_i^{EB}-\hat{\xi}_i(\boldsymbol{\theta})|\boldsymbol{y}_i]\\ &\quad= E\Big[\frac{\rho n_i}{1+\rho n_i}z_{i,\lambda}(\lambda)(\widehat{\lambda}-\lambda) + \frac{n_i}{(1+\rho n_i)^2}(\widehat{\rho}-\rho)(z_i(\lambda)-\bar{\boldsymbol{x}}_i'\boldsymbol{\beta})\Big|\boldsymbol{y}_i\Big] + O_p(m^{-1})\\ &\quad= \frac{\rho n_i}{1+\rho n_i}z_{i,\lambda}(\lambda)E[\widehat{\lambda}-\lambda|\boldsymbol{y}_i] + \frac{n_i}{(1+\rho n_i)^2}(z_i(\lambda)-\bar{\boldsymbol{x}}_i'\boldsymbol{\beta})E[\widehat{\rho}-\rho|\boldsymbol{y}_i] + O_p(m^{-1}),\end{aligned}$$

which is of order $O_p(m^{-1})$ from Lemma 4. A similar evaluation for $\widehat{\sigma}_i-\sigma_i$ shows that given y_i,

$$\widehat{\sigma}_i-\sigma_i = \frac{1}{2}\sigma_v^{-1}(1+n_i\rho)^{-1/2}(\widehat{\sigma}_v^2-\sigma_v^2) - \frac{n_i}{2}\sigma_i^{-3}(\widehat{\rho}-\rho)+O_p(m^{-1}).$$

Then, from Lemma 4, it follows that $(\widehat{\sigma}_i-\sigma_i)|\boldsymbol{y}_i = O_p(m^{-1/2})$ and $E[\widehat{\sigma}_i-\sigma_i|\boldsymbol{y}_i] = O_p(m^{-1})$, which completes the proof. □

A.6 Proof of Theorem 2. As in the proof of Theorem 1, we obtain an asymptotic expansion of $\mathcal{L}_m^c(q|\boldsymbol{y}_i)$ in the same settings of the proof of Theorem 1. Then for any $q\in\mathbb{R}$, we have

$$\mathcal{L}_m^c(q|\boldsymbol{y}_i) = E[\Phi(q+R(q,\boldsymbol{Y}))|\boldsymbol{y}_i].$$

Since $E[R(q, \boldsymbol{Y})|\boldsymbol{y}_i] = O_p(m^{-1})$, we have an asymptotic expansion of $\mathcal{L}_m^c(q|\boldsymbol{y}_i)$ as

$$\mathcal{L}_m^c(q|\boldsymbol{y}_i) = \Phi(q) + m^{-1}\eta(q, \boldsymbol{\theta}, \boldsymbol{y}_i) + O_p(m^{-3/2})$$

for an $O(1)$ smooth quantity $\eta(q, \boldsymbol{\theta}, \boldsymbol{y}_i)$, which leads to the result by Lemma 4. □

References

[1] Basu, R., Ghosh, J. K., and Mukerjee, R. (2003). Empirical Bayes prediction intervals in a normal regression model: Higher order asymptotics. *Statistics and Probability Letters*, **63**, 197-203.

[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.*, **83**, 28-36.

[3] Booth, J. G. and Hobert, J. P. (1998). Standard errors of prediction in generalized linear mixed models. *Journal of the American Statistical Association*, **93**, 262-272.

[4] Box, G.E.P. and Cox, D.R. (1964). An analysis of transformation (with discussion). *Journal of Royal Statistical Society B*, **26**, 211-252.

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

[6] Das, K., Jiang, J. and Rao, J.N.K. (2004). Mean squared error of empirical predictor. *The Annals of Statistics*, **32**, 818-840.

[7] Datta, G.S., Kubokawa, T., Molina, I., and Rao, J.N.K. (2011). Estimation of mean squared error of model-based small area estimators. *Test*, **20**, 367-388.

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

[9] 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.

[10] Fay, R. and Herriot, R. (1979). Estimators of income for small area places: an application of James–Stein procedures to census. *Journal of American Statistical Association*, **74**, 341-353.

[11] Ghosh, M. and Rao, J.N.K. (1994). Small area estimation: An appraisal. *Statist. Science*, **9**, 55-93.

[12] Hall, P. and Maiti, T. (2006a). Nonparametric estimation of mean-squared prediction error in nested-error regression models. *The Annals of Statistics*, **34**, 1733-1750.

[13] Hall, P. and Maiti, T. (2006b). On parametric bootstrap methods for small area prediction. *Journal of Royal Statistical Society*, **68**, 221-238.

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

[15] Prasad, N. and Rao, J. N. K. (1990). The estimation of mean-squared errors of small-area estimators. *Journal of American Statistical Association*, **90**, 758-766.

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

[17] Slud, E.V. and Maiti, T. (2006). Mean-squared error estimation in transformed Fay-Herriot models. *Journal of Royal Statistical Society B*, **68**, 239-257.

[18] Sugasawa, S. and Kubokawa, T. (2013). Parametric transformed Fay-Herriot model for small area estimation. Discussion Paper Series, CIRJE-F-911.

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