Japanese Journal of Statistics and Data Science (2019) 2:375–403
https://doi.org/10.1007/s42081-019-00038-y

ORIGINAL PAPER

Empirical Bayes methods in nested error regression models with skew-normal errors

Tatsuhiko Tsujino[1] · Tatsuya Kubokawa[2]

Received: 15 July 2018 / Accepted: 20 February 2019 / Published online: 4 March 2019

Abstract

The nested error regression (NER) model is a standard tool to analyze unit-level data in the field of small area estimation. Both random effects and error terms are assumed to be normally distributed in the standard NER model. However, in the case that asymmetry of distribution is observed in a given data, it is not appropriate to assume the normality. In this paper, we suggest the NER model with the error terms having skew-normal distributions. The Bayes estimator and the posterior variance are derived as simple forms. We also construct the estimators of the model-parameters based on the moment method. The resulting empirical Bayes (EB) estimator is assessed in terms of the conditional mean squared error, which can be estimated with second-order unbiasedness by parametric bootstrap methods. Through simulation and empirical studies, we compare the skew-normal model with the usual NER model and illustrate that the proposed model gives much more stable EB estimator when skewness is present.

Keywords Conditional mean squared error · Empirical Bayes estimator · Nested error regression model · Second-order approximation · Skew-normal distribution · Small area estimation

1 Introduction

Linear mixed models and their model-based estimators have been recognized as a useful method in small area estimation (SAE). Direct design-based estimates of small area means have large standard errors because small sizes of samples from

✉ Tatsuya Kubokawa
tatsuya@e.u-tokyo.ac.jp

Tatsuhiko Tsujino
co.dragon.710717@gmail.com

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

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

small areas, and the empirical best linear unbiased predictors (EBLUP) in the linear mixed models provide reliable estimates by "borrowing strength" from neighboring areas and using data of auxiliary variables. Such a model-based method for small area estimation has been studied extensively and actively from both theoretical and applied aspects. For comprehensive reviews of small area estimation, see Ghosh and Rao (1994), Pfeffermann (2013) and Rao and Molina (2015).

Among linear mixed models, Battese et al. (1988) used the nested error regression (NER) model for analyzing unit-level data in the context of SAE. This model consists of the two random variables, namely the random area effects depending on areas and the sampling error terms. Both the random variables are assumed to be normally distributed in standard NER models. While the normality assumption enables us to handle the NER models analytically, there is a growing demand for generalization of distributional assumptions to model a wider variety of characteristics of data. Asymmetry, or skewness, of distributions is one such example. When we analyze positive-valued data like income and price, we transform the data using the logarithm or the Box–Cox transformation to fit to the normality. However, the transformations may not always remove adequately the asymmetry of the data. In such situations, it is preferable to introduce an asymmetric distribution.

As a tool for relaxing the symmetry assumption, the skew-normal distribution suggested by Azzalini (1985) has been studied in the literature. In the linear mixed models, Arellano-Valle et al. (2005) considered the maximum likelihood (ML) estimation in the case where the random effect and/or the error term follow skew-normal distributions. The same model was treated by Arellano-Valle et al. (2007) from a Bayesian perspective. In the context of SAE, Ferraz and Moura (2012) dealt with the Fay–Herriot type area-level model with random effects and sampling errors having normal and skew-normal distributions, respectively, and constructed the hierarchical Bayes estimators. Diallo and Rao (2018) considered the NER model with both random effects and sampling errors having skew-normal distributions, and treated the estimation of complex parameters, such as poverty indicators, based on some optimization techniques. However, the asymptotic properties of the estimators were not investigated analytically.

In this paper, we consider the NER model with only the sampling errors having skew-normal distributions. This is different from the model treated by Diallo and Rao (2018), because the random effects in our model are normally distributed. A reason of this setup is that in our investigation, the estimation of the variance and skewness of the random effects may become unstable unless the number of areas is very large. These distributional assumptions also do not seem inappropriate in our empirical study given in Sect. 6.

In our setup of the NER model with the skew-normal errors, we derive the estimator of the model-parameters based on the moment method. The empirical Bayes (EB) estimators of parameters related to area means are analytically derived, and second-order unbiased estimators of their conditional mean squared errors (CMSE) are provided based on the parametric bootstrap. The performances of the suggested methods are examined through simulation and empirical studies.

The article is organized as follows. In Sect. 2, we provide a brief review on a skew-normal distribution and the setup of our model. Bayesian calculations of the posterior mean and variance are analytically described in Sect. 3. In Sect. 4, the estimation methods for the model-parameters estimation are suggested, and the second-order unbiased estimator of CMSE of the empirical Bayes estimator is provided. The performances of the proposed methods are investigated by simulation and empirical studies in Sects. 5 and 6 respectively. Concluding remarks are given in Sect. 7. Finally, all the technical proofs are provided in the Appendix.

2 Nested error regression models with skew-normal errors

2.1 Skew-normal distributions

We begin by explaining the skew-normal distribution which relaxes symmetry of the normal distribution. It has been widely studied since Azzalini (1985), mainly owing to the mathematical tractability. A random variable Z is said to follow a standard skew-normal distribution, denoted by $\mathcal{SN}(\lambda)$, if the density function is given by

$$f_Z(z) = 2\phi(z)\Phi(\lambda z), \quad z \in \mathbb{R},$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the probability density function (pdf) and cumulative distribution function (cdf) of the standard normal distribution $\mathcal{N}(0,1)$, respectively. Here $\lambda \in \mathbb{R}$ is a parameter which regulates the skewness of the distribution. When $\lambda = 0$, $f_Z(z)$ reduces to $\phi(z)$, so that $\mathcal{SN}(\lambda)$ includes the standard normal distribution as a special case. Figure 1 shows the density function of $\mathcal{SN}(\lambda)$ for several λ.

For practical use, consider the location-scale transformation of Z, $Y = \mu + \sigma Z$, for $\mu \in \mathbb{R}$ and $\sigma > 0$. Then, Y has a skew-normal distribution $\mathcal{SN}(\mu, \sigma^2, \lambda)$, whose density function is

$$f_Y(y) = \frac{2}{\sigma}\phi\Big(\frac{y-\mu}{\sigma}\Big)\Phi\Big(\lambda\frac{y-\mu}{\sigma}\Big), \quad y \in \mathbb{R}.$$

Following Azzalini (2013), the mean, variance and the third central moment of Y are, respectively, written as

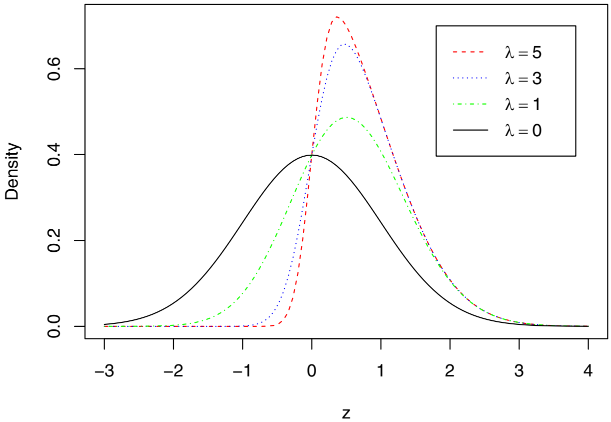

Fig. 1 The density function of $\mathcal{SN}(\lambda)$ for $\lambda = 5$ (dashed line), $\lambda = 2$ (dotted), $\lambda = 1$ (dashed dotted), $\lambda = 0$ (solid)

$$\begin{aligned} E[Y] &= \mu + \sigma\sqrt{\frac{2}{\pi}}\frac{\lambda}{\sqrt{1+\lambda^2}}, \\ m_2 = \mathrm{Var}(Y) &= \sigma^2\left(1 - \frac{2}{\pi}\frac{\lambda^2}{1+\lambda^2}\right), \end{aligned} \tag{1}$$

$$m_3 = E[(Y-E[Y])^3] = \frac{4-\pi}{2}\left(\sigma\sqrt{\frac{2}{\pi}}\frac{\lambda}{\sqrt{1+\lambda^2}}\right)^3, \tag{2}$$

which gives the skewness

$$\begin{aligned} \gamma_1 = \frac{m_3}{m_2^{3/2}} &= \frac{4-\pi}{2}\left(\sqrt{\frac{2}{\pi}}\frac{\lambda}{\sqrt{1+\lambda^2}}\right)^3\left(1 - \frac{2}{\pi}\frac{\lambda^2}{1+\lambda^2}\right)^{-3/2} \\ &= \frac{4-\pi}{2}\left(\sqrt{\frac{2}{\pi}}\delta\right)^3\left(1-\frac{2}{\pi}\delta^2\right)^{-3/2}, \end{aligned} \tag{3}$$

where $\delta = \lambda/\sqrt{1+\lambda^2}$ or $\lambda = \delta/\sqrt{1-\delta^2}$. The feasible range of γ_1 is $(-\gamma_1^{\max}, \gamma_1^{\max})$ for $\gamma_1^{\max} = \lim_{\lambda\to\infty}\gamma_1 = \lim_{\delta\to 1}\gamma_1 = \sqrt{2}(4-\pi)/(\pi-2)^{3/2} \approx 0.9953$.

To handle the skew-normal distribution analytically, the following additive representation of $Z \sim \mathcal{SN}(\lambda)$ is useful:

$$Z = \frac{1}{\sqrt{1+\lambda^2}}U_0 + \frac{\lambda}{\sqrt{1+\lambda^2}}U_1, \tag{4}$$

where U_0 and U_1 are mutually independent random variables with $U_0 \sim \mathcal{N}(0,1)$ and $U_1 \sim \mathcal{TRN}(0,1,0)$. Here $\mathcal{TRN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{d})$ denotes a p-variate truncated normal distribution with untruncated mean vector $\boldsymbol{\mu}$, untruncated covariance matrix $\boldsymbol{\Sigma}$ and the ith variable truncated below the ith element of $\boldsymbol{d}$. Omitting p implies a univariate case. For the derivation of the additive representation, see Henze (1986) and Azzalini (1986). This representation is used in the subsequent sections.

It should be remarked that there are some issues in estimation and inference of the skew-normal distribution. First, as described above, the skewness is limited in the admissible range $(-\gamma_1^{\max}, \gamma_1^{\max})$, so that skew-normal distributions cannot treat highly skewed situations. Pewsey (2000) investigated the performances of the estimators of γ_1 based on the moment method (MM) and the ML methods through a simulation study, and shows that as λ gets larger, more often the MM estimates of γ_1 fall out of the admissible range of γ_1 and the ML estimates reach the boundary values. As the sample size increases, however, the MM and ML estimates take values inside the admissible range more frequently. Another issue, as pointed out by Azzalini (1985), is that the Fisher information matrix for $Y \sim \mathcal{SN}(\mu, \sigma^2, \lambda)$ becomes singular as $\lambda \to 0$. This means that the standard asymptotic theory cannot be applicable around $\lambda = 0$. For the details, see Azzalini (1985, 2013), Azzalini and Capitanio (1999) and Pewsey (2000). Thus, in this paper, we treat the case of $\lambda \neq 0$.

2.2 Model setup and notations

In this paper, we consider m small areas, and for the ith area we have n_i observations of $(y_{ij}, \boldsymbol{x}_{ij}^\top)^\top$, $j = 1, \ldots, n_i$, where $\boldsymbol{x}_{ij} = (z_{00}, \boldsymbol{z}_{i0}^\top, \boldsymbol{z}_{ij}^\top)^\top$ is a vector of covariates which consists of the following three parts: $z_{00} = 1$ is a constant term which is common in any i and j, $\boldsymbol{z}_{i0}$ is a p_1-dimensional vector of covariates which do not depend on j, and $\boldsymbol{z}_{ij}$ is a p_2-dimensional vector of covariates which depend on i and j. Let $N = \sum_{i=1}^m n_i$ denote the total sample size. Then, we consider the following NER model with skew-normal errors:

$$\begin{aligned} y_{ij} &= \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + v_i + \varepsilon_{ij} \\ &= \beta_0 + \boldsymbol{z}_{i0}^\top \boldsymbol{\beta}_1 + \boldsymbol{z}_{ij}^\top \boldsymbol{\beta}_2 + v_i + \varepsilon_{ij}, \end{aligned} \tag{5}$$

for $j = 1, \ldots, n_i$ and $i = 1, \ldots, m$, where $\boldsymbol{\beta} = \left(\beta_0, \boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top\right)^\top$ is a $(1 + p_1 + p_2)$-dimensional unknown vector of regression coefficients. Assume that v_i's and ε_{ij}'s are mutually independent random variables with $v_i \sim \mathcal{N}(0, \tau^2)$, $\varepsilon_{ij} \sim \mathcal{SN}(0, \sigma^2, \lambda)$. From (4), ε_{ij} is expressed as:

$$\varepsilon_{ij} = \frac{\sigma}{\sqrt{1+\lambda^2}} u_{0ij} + \frac{\sigma\lambda}{\sqrt{1+\lambda^2}} u_{1ij},$$

where u_{0ij}'s and u_{1ij}'s are mutually independent, $u_{0ij} \sim \mathcal{N}(0, 1)$, and $u_{1ij} \sim \mathcal{TRN}(0, 1, 0)$. Note that we do not adjust the location of the error term to zero, so that for $\lambda \neq 0$, the error has the non-zero mean

$$\mu_\varepsilon = E[\varepsilon_{ij}] = \sigma \sqrt{\frac{2}{\pi}} \frac{\lambda}{\sqrt{1+\lambda^2}} \neq 0.$$

This implies that the constant term β_0 differs from that in the normal case.

Let $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})^\top$, $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in_i})^\top$ and $\boldsymbol{\epsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{in_i})^\top$. The model (5) is expressed in a matricial form as

$$\boldsymbol{y}_i = \boldsymbol{X}_i \boldsymbol{\beta} + v_i \boldsymbol{1}_{n_i} + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, m,$$

where $\boldsymbol{1}_n$ denotes a vector of size n with all elements equal to one. Also, $\boldsymbol{\epsilon}_i$ is written as

$$\boldsymbol{\epsilon}_i = \frac{\sigma}{\sqrt{1+\lambda^2}} \boldsymbol{u}_{0i} + \frac{\sigma\lambda}{\sqrt{1+\lambda^2}} \boldsymbol{u}_{1i},$$

where $\boldsymbol{u}_{0i} = (u_{0i1}, \dots, u_{0in_i})^\top$ and $\boldsymbol{u}_{1i} = (u_{1i1}, \dots, u_{1in_i})^\top$. Throughout the paper, we use the notations $\bar{y}_i = n_i^{-1} \sum_{j=1}^{n_i} y_{ij}$ and $\bar{\boldsymbol{x}}_i = n_i^{-1} \sum_{j=1}^{n_i} \boldsymbol{x}_{ij}$. The vector of unknown parameters is denoted by $\boldsymbol{\omega} = \left(\boldsymbol{\beta}^\top, \sigma^2, \tau^2, \lambda\right)^\top$.

3 Bayesian calculation on the predictor

3.1 Bayes estimator

We now consider the problem of estimating (predicting) $\theta_i = \boldsymbol{c}^\top\boldsymbol{\beta} + v_i$ for known $\boldsymbol{c}$. The Bayes estimator of θ_i, denoted by $\hat{\theta}_i^B(\boldsymbol{\omega})$, is given by the posterior mean of θ_i:

$$\hat{\theta}_i^B(\boldsymbol{\omega}) = E[\theta_i \,|\, \boldsymbol{y}_i] = \boldsymbol{c}^\top\boldsymbol{\beta} + E[v_i \,|\, \boldsymbol{y}_i].$$

To evaluate $E[v_i \,|\, \boldsymbol{y}_i]$, it is noted that the conditional density function of y_{ij} given v_i and u_{1ij} is given by

$$f(y_{ij} \,|\, v_i, u_{1ij}) = \phi\left(y_{ij}; \boldsymbol{x}_{ij}^\top\boldsymbol{\beta} + v_i + \frac{\sigma\lambda}{\sqrt{1+\lambda^2}} u_{1ij}, \frac{\sigma^2}{1+\lambda^2} \right),$$

where $\phi_p(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the density function of $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, a p-variate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Also, the density function of v_i and u_{1ij} are respectively $f(v_i) = \phi(v_i; 0, \tau^2)$ and $f(u_{1ij}) = 2\phi(u_{1ij})I(u_{1ij} > 0)$

where $I(\cdot)$ is an indicator function. The conditional density function of $(v_i, \boldsymbol{u}_{1i})$ given $\boldsymbol{y}_i$ is written as

$$f(v_i, \boldsymbol{u}_{1i} \,|\, \boldsymbol{y}_i) \propto \left\{ \prod_{j=1}^{n_i} f(y_{ij} \,|\, v_i, u_{1ij}) f(u_{1ij}) \right\} f(v_i).$$

To rewrite the conditional density, let

$$\begin{aligned} \mu_{v_i} &= \frac{n_i\tau^2(1+\lambda^2)}{\sigma^2 + n_i\tau^2(1+\lambda^2)} \left(\bar{y}_i - \bar{\boldsymbol{x}}_i^\top\boldsymbol{\beta} - \frac{\sigma\lambda}{\sqrt{1+\lambda^2}} n_i^{-1} \sum_{j=1}^{n_i} u_{1ij} \right), \\ \sigma_{v_i}^2 &= \frac{\sigma^2\tau^2}{\sigma^2 + n_i\tau^2(1+\lambda^2)}. \end{aligned}$$

Also, let $\boldsymbol{R}_i = (1-\rho_i)\boldsymbol{I}_{n_i} + \rho_i \boldsymbol{1}_{n_i}\boldsymbol{1}_{n_i}^\top$ for the $n \times n$ identity matrix $\boldsymbol{I}_n$ and

$$\rho_i = \frac{\tau^2\lambda^2/(\sigma^2 + n_i\tau^2)}{1 + \tau^2\lambda^2/(\sigma^2 + n_i\tau^2)}.$$

Denote the (j, k) element of $\boldsymbol{R}_i$ by $\rho_{i,jk}$, namely,

$$\rho_{i,jk} = \begin{cases} 1 & \text{for } j = k, \\ \rho_i & \text{for } j \neq k. \end{cases}$$

Then, the conditional density can be rewritten as

$$f(v_i, \boldsymbol{u}_{1i} \,|\, \boldsymbol{y}_i) \propto \phi(v_i; \mu_{v_i}, \sigma_{v_i}^2) \phi_{n_i}(\boldsymbol{u}_{1i}; \boldsymbol{\mu}_i, \sigma_{u_i}^2 \boldsymbol{R}_i) \prod_{j=1}^{n_i} I(u_{1ij} > 0), \tag{6}$$

where $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{in_i})^\top$ for

$$\mu_{ij} = \frac{\lambda}{\sigma\sqrt{1+\lambda^2}} \left\{ y_{ij} - \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} - \frac{n_i \tau^2}{\sigma^2 + n_i \tau^2} (\bar{y}_i - \bar{\boldsymbol{x}}_i^\top \boldsymbol{\beta}) \right\},$$

and

$$\sigma_{u_i}^2 = \frac{1}{1+\lambda^2} \left(1 + \frac{\tau^2 \lambda^2}{\sigma^2 + n_i \tau^2} \right).$$

The derivation of (6) is given in the Appendix. Let

$$\boldsymbol{w}_i = (w_{i1}, \ldots, w_{in_i})^\top = (\boldsymbol{u}_{1i} - \boldsymbol{\mu}_i)/\sigma_{u_i} \quad \text{and} \quad \boldsymbol{a}_i = \boldsymbol{\mu}_i / \sigma_{u_i}.$$

Then, it can be seen that $\boldsymbol{w}_i \,|\, \boldsymbol{y}_i \sim \mathcal{TRN}_{n_i}(\boldsymbol{0}, \boldsymbol{R}_i, -\boldsymbol{a}_i)$ and

$$\begin{aligned} E[v_i \,|\, \boldsymbol{y}_i] &= E[E[v_i \,|\, \boldsymbol{y}_i, \boldsymbol{u}_{1i}] \,|\, \boldsymbol{y}_i] \\ &= \frac{n_i \tau^2}{\sigma^2 + n_i \tau^2} (\bar{y}_i - \bar{\boldsymbol{x}}_i^\top \boldsymbol{\beta}) - \frac{\sigma\lambda}{\sqrt{1+\lambda^2}} \frac{n_i \tau^2 (1+\lambda^2)}{\sigma^2 + n_i \tau^2 (1+\lambda^2)} \sigma_{u_i} n_i^{-1} \sum_{j=1}^{n_i} E[w_{ij} \,|\, \boldsymbol{y}_i]. \end{aligned} \tag{7}$$

In the last expression, we need to calculate the conditional mean of w_{ij} given $\boldsymbol{y}_i$. As shown in Tallis (1961), the moment of a multivariate truncated normal distribution involves multiple integrals. However, Diallo and Rao (2018) pointed out that a simple structure of $\boldsymbol{R}_i$ allows us to reduce these multiple integrals to one-dimensional integrals on a product of univariate normal distribution's cdf's. This simplification leads to a significant reduction of computational cost and gives a clear expression of $E[w_{ij} \,|\, \boldsymbol{y}_i]$ described in the following lemma. All the proofs in this section are provided in the Appendix.

Lemma 3.1 *For* $j = 1, \ldots, n_i$,

$$E[w_{ij} \,|\, \boldsymbol{y}_i] = \sum_{k=1}^{n_i} \rho_{i,jk} \phi(a_{ik}) \frac{\alpha_{1ik}}{\alpha_{0i}},$$

where

$$\alpha_{0i} = \int_{-\infty}^{\infty} \left\{ \prod_{j=1}^{n_i} \phi\left(\frac{a_{ij} + \sqrt{\rho_i}\xi}{\sqrt{1-\rho_i}} \right) \right\} \phi(\xi)\,\mathrm{d}\xi, \tag{8}$$

$$\alpha_{1ij} = \int_{-\infty}^{\infty} \left\{ \prod_{k \neq j} \Phi\left(\frac{a_{ik} - \rho_i a_{ij}}{\sqrt{1-\rho_i}} + \sqrt{\rho_i}\xi \right) \right\} \phi(\xi)\,\mathrm{d}\xi. \tag{9}$$

Using this result, we derive the Bayes estimator $\hat{\theta}_i^B(\boldsymbol{\omega})$ as presented in the following theorem.

Theorem 3.1 *The Bayes estimator of* θ_i *is given by*

$$\hat{\theta}_i^B(\boldsymbol{\omega}) = \boldsymbol{c}^\top \boldsymbol{\beta} + \frac{n_i\tau^2}{\sigma^2 + n_i\tau^2} \left\{ \bar{y}_i - \bar{\boldsymbol{x}}_i^\top \boldsymbol{\beta} - \frac{\sigma\lambda}{\sqrt{1+\lambda^2}} n_i^{-1} \sum_{j=1}^{n_i} \sigma_{u_i}^{-1} \phi(a_{ij}) \frac{\alpha_{1ij}}{\alpha_{0i}} \right\}. \tag{10}$$

It is noted that $\boldsymbol{c}^\top\boldsymbol{\beta} + n_i\tau^2(\sigma^2 + n_i\tau^2)^{-1}(\bar{y}_i - \bar{\boldsymbol{x}}_i^\top\boldsymbol{\beta})$ corresponds to the Bayes estimator in the NER model under normality. Thus the last term in the parenthesis is interpreted as a correction term for the mean μ_ε of the skew-normal error.

3.2 Posterior variance

We next calculate the posterior variance of θ_i. Since $\theta_i - \hat{\theta}_i^B(\boldsymbol{\omega}) = v_i - E[v_i \,|\, \boldsymbol{y}_i]$, we have

$$\begin{aligned} \mathrm{Var}(\theta_i \,|\, \boldsymbol{y}_i) &= \mathrm{Var}(v_i \,|\, \boldsymbol{y}_i) \\ &= E[\sigma_{v_i}^2 \,|\, \boldsymbol{y}_i] + \mathrm{Var}(\mu_{v_i} \,|\, \boldsymbol{y}_i) \\ &= \frac{\sigma^2\tau^2}{\sigma^2 + n_i\tau^2(1+\lambda^2)} + \frac{\sigma^2\tau^4\lambda^2(1+\lambda^2)}{\{\sigma^2 + n_i\tau^2(1+\lambda^2)\}^2} \sigma_{u_i}^2 \mathrm{Var}\left(\sum_{j=1}^{n_i} w_{ij} \,\middle|\, \boldsymbol{y}_i \right), \end{aligned} \tag{11}$$

so that we need to calculate the second moments of $\boldsymbol{w}_i$ given $\boldsymbol{y}_i$. Using the special structure of $\boldsymbol{R}_i$, we obtain an easy-to-calculate expression of $E[w_{ij}w_{ik} \,|\, \boldsymbol{y}_i]$.

Lemma 3.2 *For any* $j, k = 1, \dots, n_i$,

$$\begin{aligned} E[w_{ij}w_{ik} \,|\, \boldsymbol{y}_i] = \rho_{i,jk} &- \sum_{q=1}^{n_i} \rho_{i,jq}\rho_{i,kq} a_{iq} \phi(a_{iq}) \frac{\alpha_{1iq}}{\alpha_{0i}} \\ &+ \sum_{q=1}^{n_i} \rho_{i,jq} \sum_{r \neq q} (\rho_{i,kr} - \rho_i \rho_{i,kq}) \phi(a_{iq}, a_{ir}; \rho_i) \frac{\alpha_{2iqr}}{\alpha_{0i}}, \end{aligned}$$

where $\phi(\cdot,\cdot;\rho)$ is the density function of $\mathcal{N}_2(\mathbf{0},\boldsymbol{R})$ for the correlation matrix $\boldsymbol{R}$ with off-diagonal elements ρ, and

$$\alpha_{2iqr} = \int_{-\infty}^{\infty} \left\{ \prod_{s\neq q,r} \Phi\left(\frac{a_{is} - \rho_i(1+\rho_i)^{-1}(a_{iq}+a_{ir})}{\sqrt{1-\rho_i}} + \sqrt{\frac{\rho_i}{1+\rho_i}}\xi \right) \right\} \phi(\xi)\,\mathrm{d}\xi. \tag{12}$$

Using Lemma 3.2, we obtain a tractable form of the posterior variance.

Theorem 3.2 *The posterior variance of θ_i given $\boldsymbol{y}_i$ is*

$$\mathrm{Var}(\theta_i \mid \boldsymbol{y}_i) = \frac{\sigma^2\tau^2}{\sigma^2 + n_i\tau^2} - \rho_i v_i(\boldsymbol{\omega},\boldsymbol{y}_i)\frac{\sigma^2\tau^2}{\sigma^2+n_i\tau^2}, \tag{13}$$

where

$$\begin{aligned} v_i(\boldsymbol{\omega},\boldsymbol{y}_i) = & \sum_{j=1}^{n_i} a_{ij}\phi(a_{ij})\frac{\alpha_{1ij}}{\alpha_{0i}} \\ & - (1-\rho_i)\sum_{j=1}^{n_i}\sum_{k\neq j}\phi(a_{ij},a_{ik};\rho_i)\frac{\alpha_{2ijk}}{\alpha_{0i}} + \left\{ \sum_{j=1}^{n_i} \phi(a_{ij})\frac{\alpha_{1ij}}{\alpha_{0i}} \right\}^2. \end{aligned} \tag{14}$$

It is noted that the first term of RHS in (13) is identical to the posterior variance of θ_i in the normal case, and the second term comes from the skewness of the error term.

4 Empirical Bayes (EB) estimator

4.1 Estimation of parameters

Since the Bayes estimator (10) depends on the unknown parameters $\boldsymbol{\omega} = \left(\beta_0, \boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top, \sigma^2, \tau^2, \lambda\right)^\top$, we need to estimate them from the given observations. The procedure of estimation is quite similar to the method proposed by Fuller and Battese (1973) with some adjustments made for estimating β_0, σ^2 and λ.

We first consider the estimation of $\boldsymbol{\beta}$ in the case that σ^2, τ^2 and λ are known. As described in Sect. 2.2, since the location of the error term is not centered at zero in our setup, it is seen that

$$E[y_{ij}] = \boldsymbol{x}_{ij}^\top\boldsymbol{\beta} + \mu_\varepsilon = \beta_{0\varepsilon} + \boldsymbol{z}_{i0}^\top\boldsymbol{\beta}_1 + \boldsymbol{z}_{ij}^\top\boldsymbol{\beta}_2 = \boldsymbol{x}_{ij}^\top\boldsymbol{\beta}_\varepsilon,$$

where $\beta_{0\varepsilon} = \beta_0 + \mu_\varepsilon$. The covariance matrix of $\boldsymbol{y}_i$ is given by

$$\boldsymbol{V}_i = \boldsymbol{V}_i(\sigma^2,\tau^2,\lambda) = m_2\boldsymbol{I}_{n_i} + \tau^2\mathbf{1}_{n_i}\mathbf{1}_{n_i}^\top.$$

for m_2 defined in (1). Assuming that the matrix $\left(\boldsymbol{X}_1^\top, \dots, \boldsymbol{X}_m^\top\right)^\top$ is of full rank, one can obtain the best linear unbiased estimator (BLUE) of $\boldsymbol{\beta}_\varepsilon$ as

$$\widetilde{\boldsymbol{\beta}}_\varepsilon = \left(\tilde{\beta}_{0\varepsilon}, \widetilde{\boldsymbol{\beta}}_1^\top, \widetilde{\boldsymbol{\beta}}_2^\top\right)^\top = \left(\sum_{i=1}^m \boldsymbol{X}_i^\top \boldsymbol{V}_i^{-1} \boldsymbol{X}_i\right)^{-1} \sum_{i=1}^m \boldsymbol{X}_i^\top \boldsymbol{V}_i^{-1} \boldsymbol{y}_i.$$

Then, the estimator $\widetilde{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is obtained as $\widetilde{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\beta}}(\sigma^2, \tau^2, \lambda) = \left(\tilde{\beta}_0, \widetilde{\boldsymbol{\beta}}_1^\top, \widetilde{\boldsymbol{\beta}}_2^\top\right)^\top$ where

$$\tilde{\beta}_0 = \tilde{\beta}_{0\varepsilon} - \mu_\varepsilon = \tilde{\beta}_{0\varepsilon} - \sigma\sqrt{\frac{2}{\pi}}\frac{\lambda}{\sqrt{1+\lambda^2}}.$$

In practice, the parameters $(\sigma^2, \tau^2, \lambda)$ are unknown, so that we need to estimate them beforehand. For those areas with $n_i \geq 2$, we take the deviations from the group mean $\overline{y}_i = n_i^{-1}\sum_{j=1}^{n_i} y_{ij}$ for the model (5), which gives

$$\tilde{y}_{ij} = \tilde{\boldsymbol{z}}_{ij}^\top \boldsymbol{\beta}_2 + \tilde{\varepsilon}_{ij},$$

where $\tilde{y}_{ij} = y_{ij} - \overline{y}_i$, $\tilde{\boldsymbol{z}}_{ij} = \boldsymbol{z}_{ij} - \overline{\boldsymbol{z}}_i$, $\tilde{\varepsilon}_{ij} = \varepsilon_{ij} - \overline{\varepsilon}_i$, and $\overline{\boldsymbol{z}}_i$ and $\overline{\varepsilon}_i$ are defined analogously to $\overline{y}_i$. In the method of Fuller and Battese (1973), it is sufficient to consider the second moment of $\tilde{\varepsilon}_{ij}$ for estimating the variance parameter σ^2. However, we use the second and third moment to jointly estimate σ^2 and the skewness parameter λ. Here we see that for $j = 1, \dots, n_i$,

$$\begin{aligned} E[\tilde{\varepsilon}_{ij}^2] &= E[\{(\varepsilon_{ij} - \mu_\varepsilon) - (\overline{\varepsilon}_i - \mu_\varepsilon)\}^2] \\ &= E[(\varepsilon_{ij} - \mu_\varepsilon)^2] - 2n_i^{-1}E[(\varepsilon_{ij} - \mu_\varepsilon)^2] + n_i^{-1}E[(\varepsilon_{ij} - \mu_\varepsilon)^2] \\ &= \{(n_i - 1)/n_i\}m_2, \end{aligned} \tag{15}$$

$$\begin{aligned} E[\tilde{\varepsilon}_{ij}^3] &= E[\{(\varepsilon_{ij} - \mu_\varepsilon) - (\overline{\varepsilon}_i - \mu_\varepsilon)\}^3] \\ &= E[(\varepsilon_{ij} - \mu_\varepsilon)^3] - 3n_i^{-1}E[(\varepsilon_{ij} - \mu_\varepsilon)^3] + 3n_i^{-2}E[(\varepsilon_{ij} - \mu_\varepsilon)^3] - n_i^{-2}E[(\varepsilon_{ij} - \mu_\varepsilon)^3] \\ &= \{(n_i - 1)(n_i - 2)/n_i^2\}m_3, \end{aligned} \tag{16}$$

for m_3 defined in (1). It is noted that the third moment can be used for those areas with $n_i \geq 3$. Assume that $(\tilde{\boldsymbol{z}}_{11}, \dots, \tilde{\boldsymbol{z}}_{1n_1}, \dots, \tilde{\boldsymbol{z}}_{m1}, \dots, \tilde{\boldsymbol{z}}_{mn_m})^\top$ has full column rank and let $\hat{\varepsilon}_{ij}$ be the residual obtained by regressing $\tilde{y}_{ij}$ on $\tilde{\boldsymbol{z}}_{ij}$. Since $\delta = \lambda/\sqrt{1+\lambda^2}$, based on (15) and (16) we get estimators of σ^2 and δ as the solution of the following equations:

$$\begin{aligned} \widehat{m}_2 &= \frac{\mathrm{RSS}(1)}{N - m - p_2} = \sigma^2\left(1 - \frac{2}{\pi}\delta^2\right) = m_2, \\ \widehat{m}_3 &= \frac{\mathrm{RSC}(1)}{\eta_1} = \frac{4-\pi}{2}\left(\sigma\sqrt{\frac{2}{\pi}}\delta\right)^3 = m_3, \end{aligned} \tag{17}$$

where $\eta_1 = \sum_{i=1}^m n_i^{-1}(n_i-1)(n_i-2)$, $\mathrm{RSS}(1) = \sum_{i=1}^m \sum_{j=1}^{n_i} \hat{\varepsilon}_{ij}^2$ is the residual sum of squares and $\mathrm{RSC}(1) = \sum_{i=1}^m \sum_{j=1}^{n_i} \hat{\varepsilon}_{ij}^3$ is the residual sum of cubes. Note that the degrees of freedom associated with the residual is taken into account only in the first formula of (17), so that $\widehat{m}_2$ is an unbiased estimator of m_2 but $\widehat{m}_3$ is not unbiased. Solving the equations (17) gives the estimators as

$$\hat{\sigma}^2 = \widehat{m}_2 + \left(\frac{2}{4-\pi}\widehat{m}_3\right)^{2/3},$$
$$\widetilde{\delta} = \frac{1}{\hat{\sigma}}\sqrt{\frac{\pi}{2}}\left(\frac{2}{4-\pi}\widehat{m}_3\right)^{1/3}.$$

We remark that $\left|\widetilde{\delta}\right| < 1$ is assumed here. As mentioned in Sect. 2.1, however, smaller sample size or larger $|\lambda|$ value may increase the possibility of violating this condition. In this case, we suggest the truncated estimators

$$\widehat{\delta} = \begin{cases} 1-1/m & \text{for } \widetilde{\delta} > 1-1/m, \\ \widetilde{\delta} & \text{for } |\widetilde{\delta}| \le 1-1/m, \\ -1+1/m & \text{for } \widetilde{\delta} < -1+1/m, \end{cases} \quad \text{and} \quad \widehat{\lambda} = \widehat{\delta}/\sqrt{1-\widehat{\delta}^2}.$$

This problem is examined in the simulation study presented in Sect. 5.

Next, using $\widehat{m}_2$, an unbiased estimator $\widetilde{\tau}^2$ of τ^2 is given by

$$\widetilde{\tau}^2 = \eta_2^{-1}\left\{\mathrm{RSS}(2) - (N-1-p_1-p_2)\widehat{m}_2\right\},$$

where RSS(2) is the residual sum of squares obtained by regressing y_{ij} on $\boldsymbol{x}_{ij}$, and

$$\eta_2 = \sum_{i=1}^m n_i\left\{1 - n_i\bar{\boldsymbol{x}}_i^\top\left(\sum_{i=1}^m \boldsymbol{X}_i^\top\boldsymbol{X}_i\right)^{-1}\bar{\boldsymbol{x}}_i\right\}.$$

Since $\widetilde{\tau}^2$ can take a negative value, we truncate $\widetilde{\tau}^2$ at zero and get the truncated estimator as $\hat{\tau}^2 = \max(\widetilde{\tau}^2, 0)$. Now $\boldsymbol{\beta}$ can be estimated as $\widehat{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\beta}}(\hat{\sigma}^2, \hat{\tau}^2, \widehat{\lambda})$. Substituting the estimator $\widehat{\boldsymbol{\omega}} = \left(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}}_1^\top, \widehat{\boldsymbol{\beta}}_2^\top, \hat{\sigma}^2, \hat{\tau}^2, \widehat{\lambda}\right)^\top$ into the Bayes estimator $\hat{\theta}_i^B(\boldsymbol{\omega})$, one gets the empirical Bayes (EB) estimator $\hat{\theta}_i^{EB} = \hat{\theta}_i^B(\widehat{\boldsymbol{\omega}})$ of θ_i.

Finally, we show the consistency and other asymptotic properties of the estimator $\widehat{\boldsymbol{\omega}}$, which implies that the EB estimator $\hat{\theta}_i^B(\widehat{\boldsymbol{\omega}})$ converges to the Bayes estimator $\hat{\theta}_i^B(\boldsymbol{\omega})$ as $m \to \infty$. The asymptotic properties are used in the next section for deriving a second-order unbiased estimator of the conditional mean squared error of $\hat{\theta}_i^{EB}$. To this end, we assume the following regularity condition:

(RC) n_i's are bounded below and above, that is, there exist positive constants $\underline{n}$ and $\bar{n}$ satisfying $\underline{n} \le n_i \le \bar{n}$. Elements of $\boldsymbol{X}_i$ are uniformly bounded, $\sum_{i=1}^m \boldsymbol{X}_i^\top \boldsymbol{V}_i^{-1}\boldsymbol{X}_i$ is a positive definite matrix, and $m^{-1}\sum_{i=1}^m \boldsymbol{X}_i^\top \boldsymbol{V}_i^{-1}\boldsymbol{X}_i$ converges to a positive definite matrix.

As remarked before, the standard asymptotic theory does not hold when $\lambda = 0$. Thus, in addition to (RC), the condition $\lambda \neq 0$ is required to obtain the usual rate of convergence in the skew-normal case.

Theorem 4.1 *If* (RC) *and* $\lambda \neq 0$ *hold, then*

$$\begin{aligned} E\big[(\hat{\boldsymbol{\omega}} - \boldsymbol{\omega})(\hat{\boldsymbol{\omega}} - \boldsymbol{\omega})^\top \mid \boldsymbol{y}_i\big] &= O_p(m^{-1}), \\ E[\hat{\boldsymbol{\omega}} - \boldsymbol{\omega} \mid \boldsymbol{y}_i] &= O_p(m^{-1}). \end{aligned} \tag{18}$$

Thus it follows from (18) *that* $\hat{\boldsymbol{\omega}} - \boldsymbol{\omega} \mid \boldsymbol{y}_i = O_p(m^{-1/2})$.

4.2 Measuring uncertainty of the EB estimator

In this section, we evaluate the conditional mean squared error of the EB estimator for measuring the uncertainty. Although the unconditional mean squared error (MSE) is often used as a measure of uncertainty of the predictors, we employ the conditional mean squared error (CMSE) because the area-specific prediction of θ_i given $\boldsymbol{y}_i$ is our main interest. The CMSE is initially proposed by Booth and Hobert (1998), suggesting that the MSE is inappropriate when small domains are supposed in mixed model settings and researchers focus on area-specific prediction. The CMSE of the EB estimator is defined as

$$\mathrm{CMSE}_i(\boldsymbol{\omega}, \boldsymbol{y}_i) = E\big[\big\{\hat{\theta}_i^B(\hat{\boldsymbol{\omega}}) - \theta_i\big\}^2 \mid \boldsymbol{y}_i\big],$$

which is decomposed as

$$\mathrm{CMSE}_i(\boldsymbol{\omega}, \boldsymbol{y}_i) = E\big[\big\{\hat{\theta}_i^B(\boldsymbol{\omega}) - \theta_i\big\}^2 \mid \boldsymbol{y}_i\big] + E\big[\big\{\hat{\theta}_i^B(\hat{\boldsymbol{\omega}}) - \hat{\theta}_i^B(\boldsymbol{\omega})\big\}^2 \mid \boldsymbol{y}_i\big],$$

because the cross product term is $E[E[\{\hat{\theta}_i^B(\omega) - \theta_i\}\{\hat{\theta}_i^B(\hat{\omega}) - \hat{\theta}_i^B(\omega)\} \mid \boldsymbol{y}_1, \dots, \boldsymbol{y}_m] \mid \boldsymbol{y}_i] = E[\{\hat{\theta}_i^B(\omega) - E[\theta_i \mid \boldsymbol{y}_1, \dots, \boldsymbol{y}_m]\} \{\hat{\theta}_i^B(\hat{\omega}) - \hat{\theta}_i^B(\omega)\} \mid \boldsymbol{y}_i = 0$. Let $g_{1i}(\boldsymbol{\omega}, \boldsymbol{y}_i) = E\big[\big\{\hat{\theta}_i^B(\boldsymbol{\omega}) - \theta_i\big\}^2 \mid \boldsymbol{y}_i\big]$ and $g_{2i}(\boldsymbol{\omega}, \boldsymbol{y}_i) = E\big[\big\{\hat{\theta}_i^B(\hat{\boldsymbol{\omega}}) - \hat{\theta}_i^B(\boldsymbol{\omega})\big\}^2 \mid \boldsymbol{y}_i\big]$. It is seen that $g_{1i}(\boldsymbol{\omega}, \boldsymbol{y}_i)$ is the CMSE of the Bayes estimator and $g_{2i}(\boldsymbol{\omega}, \boldsymbol{y}_i)$ involves the uncertainty on the estimation of the model parameters. Since the CMSE of the Bayes estimator corresponds to the posterior variance, from (13) in Theorem 3.2, we have

$$g_{1i}(\boldsymbol{\omega}, \boldsymbol{y}_i) = \mathrm{Var}(\theta_i \mid \boldsymbol{y}_i) = \{\sigma^2\tau^2/(\sigma^2 + n_i\tau^2)\}\{1 - \rho_i v_i(\boldsymbol{\omega}, \boldsymbol{y}_i)\}.$$

Also, $g_{2i}(\boldsymbol{\omega}, \boldsymbol{y}_i)$ can be approximated as

$$g_{2i}(\boldsymbol{\omega}, \boldsymbol{y}_i) = E\left[\left\{(\hat{\boldsymbol{\omega}} - \boldsymbol{\omega})^\top \frac{\partial}{\partial \boldsymbol{\omega}} \hat{\theta}_i^B(\boldsymbol{\omega})\right\}^2 \middle| \boldsymbol{y}_i\right] + o_p(m^{-1}),$$

which implies that $g_{2i}(\boldsymbol{\omega}, \boldsymbol{y}_i) = O_p(m^{-1})$.

We derive the second-order unbiased estimator of the CMSE for $\lambda \neq 0$ case. Since it is difficult to calculate analytically the second-order approximations of $g_{1i}(\boldsymbol{\omega}, \boldsymbol{y}_i)$ and $g_{2i}(\boldsymbol{\omega}, \boldsymbol{y}_i)$, we use parametric bootstrap methods. Consider conditioning on the area i, thus $\boldsymbol{y}_i$ is fixed. A parametric bootstrap sample $\boldsymbol{y}_k^* = (y_{k1}^*, \dots, y_{kn_k}^*)^\top$ is generated from the model

$$\begin{aligned} y_{kj}^* &= \boldsymbol{x}_{kj}^\top \widehat{\boldsymbol{\beta}} + v_k^* + \varepsilon_{kj}^*, \quad j = 1, \dots, n_k,\ k = 1, \dots, m,\ k \neq i, \\ v_k^* &\sim \mathcal{N}(0, \hat{\tau}^2), \\ \varepsilon_{kj}^* &\sim \mathcal{SN}(0, \hat{\sigma}^2, \widehat{\lambda}), \end{aligned}$$

where v_i^*'s and ε_{ij}^*'s are mutually independent. We construct the estimator $\widehat{\boldsymbol{\omega}}^*$ using the same method as used to obtain $\widehat{\boldsymbol{\omega}}$ except that the bootstrap sample

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

is used to obtain $\widehat{\boldsymbol{\omega}}^*$. It is noted that the given data $\boldsymbol{y}_i$ is used in the bootstrap sample, because we consider the conditional expectation given $\boldsymbol{y}_i$. Since we have the analytical form of $g_{1i}(\boldsymbol{\omega}, \boldsymbol{y}_i)$ as (13), a second-order unbiased estimator of $g_{1i}(\boldsymbol{\omega}, \boldsymbol{y}_i)$ is given by

$$\hat{g}_{1i} = 2g_{1i}(\widehat{\boldsymbol{\omega}}, \boldsymbol{y}_i) - E^*[g_{1i}(\widehat{\boldsymbol{\omega}}^*, \boldsymbol{y}_i) \mid \boldsymbol{y}_i],$$

where $E^*[\cdot \mid \boldsymbol{y}_i]$ is the expectation with respect to the bootstrap sample. Also, $g_{2i}(\boldsymbol{\omega}, \boldsymbol{y}_i)$ is estimated by

$$\hat{g}_{2i} = E^*\left[\left\{\hat{\theta}_i^B(\widehat{\boldsymbol{\omega}}^*) - \hat{\theta}_i^B(\widehat{\boldsymbol{\omega}})\right\}^2 \mid \boldsymbol{y}_i\right].$$

It can be shown that $E[\hat{g}_{1i} \mid \boldsymbol{y}_i] = g_{1i}(\boldsymbol{\omega}, \boldsymbol{y}_i) + o_p(m^{-1})$ and $E[\hat{g}_{2i} \mid \boldsymbol{y}_i] = g_{2i}(\boldsymbol{\omega}, \boldsymbol{y}_i) + o_p(m^{-1})$, thus a second-order unbiased estimator of the CMSE is given by

$$\widehat{\mathrm{CMSE}}_i = \hat{g}_{1i} + \hat{g}_{2i}.$$

Proposition 4.1 *Assume* (RC) *and* $\lambda \neq 0$. *Then,* $\widehat{\mathrm{CMSE}}_i$ *is a second-order unbiased estimator of the CMSE*:

$$E\left[\widehat{\mathrm{CMSE}}_i \mid \boldsymbol{y}_i\right] = \mathrm{CMSE}_i(\boldsymbol{\omega}, \boldsymbol{y}_i) + o_p(m^{-1}).$$

Table 1 Biases and square roots of MSEs for $\left(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2, \hat{\sigma}^2, \hat{\tau}^2, \widehat{\lambda}\right)$ and PFE

λ	m	n	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\sigma}^2$	$\hat{\tau}^2$	$\hat{\lambda}$	PFE
1	20	5	34.196 (79.14)	− 0.407 (33.88)	− 0.042 (8.37)	0.169 (0.38)	− 0.007 (0.37)	− 0.461 (1.60)	99.2
		10	22.894 (62.16)	− 0.640 (33.19)	− 0.192 (6.41)	0.058 (0.25)	− 0.003 (0.36)	− 0.311 (1.10)	100.0
	70	5	21.064 (54.83)	− 0.058 (13.47)	− 0.095 (5.05)	0.045 (0.22)	− 0.004 (0.20)	− 0.286 (1.03)	100.0
		10	10.269 (36.31)	− 0.064 (12.98)	0.000 (3.44)	− 0.002 (0.16)	− 0.002 (0.18)	− 0.144 (0.63)	100.0
6	20	5	6.933 (30.96)	− 0.404 (32.91)	− 0.014 (6.27)	− 0.036 (0.29)	− 0.053 (0.36)	− 3.307 (3.37)	70.7
		10	3.804 (26.85)	− 0.684 (32.72)	− 0.099 (4.76)	− 0.016 (0.19)	− 0.052 (0.35)	− 3.074 (3.09)	74.2
	70	5	1.736 (14.31)	− 0.147 (13.10)	− 0.051 (3.83)	− 0.011 (0.16)	− 0.017 (0.19)	− 1.347 (1.88)	74.0
		10	0.743 (12.75)	− 0.031 (12.83)	− 0.007 (2.59)	− 0.005 (0.10)	− 0.011 (0.18)	− 0.998 (1.41)	82.4

For each choice of (λ, m, n) the biases are given with the square roots of MSEs provided in the parentheses below. Values for $\widehat{\beta}_0, \widehat{\beta}_1$ and $\widehat{\beta}_2$ are multiplied by 100

5 Simulation study

In this section we examine the performance of the proposed methods numerically through Monte Carlo simulation. Through this section, we consider the simple nested error regression model $y_{ij} = \beta_0 + z_{i0}\beta_1 + z_{ij}\beta_2 + v_i + \varepsilon_{ij}$ for $j = 1, \dots, n$, $i = 1, \dots, m$, where values of z_{i0} and z_{ij} are generated from $\mathcal{N}(0, 1)$ and fixed throughout the simulation runs. For simplicity, we set $\beta_0 = \beta_1 = \beta_2 = \sigma^2 = \tau^2 = 1$ and conduct the simulation experiment for different values of m, n and λ.

We first investigate the performance of the suggested estimators of the model-parameters. For the parameter λ we treat the two cases of $\lambda = 1$ and $\lambda = 6$ which correspond to $\gamma_1 = 0.3873$ and $\gamma_1 = 0.9285$ for γ_1 defined in (3). Concerning (m, n), we treat four combinations $(m, n) = (20, 5), (20, 10), (70, 5), (70, 10)$. Based on $R = 5,000$ simulation runs, we compute the bias and MSE for the estimator of each parameter, and the value of $\Pr\left\{|\widetilde{\delta}| < 1\right\}$, the probability that the estimated skewness is in the feasible range introduced in Sect. 2.1. These quantities are respectively calculated as

$$\text{Bias} = \frac{1}{R}\sum_{r=1}^{R}(r\text{th estimate} - \text{true parameter}),$$

$$\text{MSE} = \frac{1}{R}\sum_{r=1}^{R}(r\text{th estimate} - \text{true parameter})^2,$$

$$\text{PFE }(\%) = 100 \times \Pr\left\{|\widetilde{\delta}| < 1\right\} = 100 \times \frac{\#\{|r\text{th value of } \widetilde{\delta}| < 1\}}{R},$$

and those values are reported in Table 1. Note that values for $\widehat{\beta}_0$, $\widehat{\beta}_1$ and $\widehat{\beta}_2$ are multiplied by 100. It is observed that MSEs decrease as m increases, which coincides with the consistency of the estimators. As remarked in Sects. 2.1 and 4.1, it is seen that PFE increases as m and n increase, or $|\lambda|$ decreases. From Table 1, it is seen that the estimator $\widehat{\lambda}$ has large values of the bias and the MSE for $m = 20$, but it gets more stable for $m = 70$. This shows that we need data with large m to estimate λ precisely.

In the next study, we examine the performance of the estimator of the CMSE given in Sect. 4.2. Without loss of generality, we treat the prediction of $\theta_1 = \overline{\boldsymbol{x}}_1^\top \boldsymbol{\beta} + v_1$ which is related to the area mean in the area $i = 1$. The model setup is the same as in the previous simulation except that we set $n_1 = \cdots = n_m = n = 5$ throughout the simulation runs. For values of λ and m, consider the four cases $(\lambda, m) = (0, 20)$, $(0, 50)$, $(3, 20)$, $(3, 50)$. As conditioning values of y_{1j}'s, we use the qth quantile of the marginal distribution of y_{1j}, denoted by $y_{1j(q)}$, for $q = 0.05$, 0.25, 0.50, 0.75, and 0.95. $y_{1j(q)}$ is given by $y_{1j(q)} = \beta_0 + z_{10}\beta_1 + z_{1j}\beta_2 + r_{1j(q)}$ where $r_{1j(q)}$ is the qth quantile of $r_{1j} = v_1 + \varepsilon_{1j}$. Here r_{1j} is distributed as $\mathcal{SN}(0, \sigma^2 + \tau^2, \widetilde{\lambda})$ where $\widetilde{\lambda} = \sigma\lambda/\sqrt{\sigma^2 + (1 + \lambda^2)\tau^2}$. In advance, we calculate the true value of the CMSE for each q value based on $R = 100{,}000$ simulation runs as follows:

Table 2 True values of the CMSE, means and relative biases of the estimates for the CMSE

λ	m	q	$r_{1j(q)}$	SN			Normal		
				CMSE_1	$E[\widehat{\mathrm{CMSE}}_1]$	RB	CMSE_1	$E[\widehat{\mathrm{CMSE}}_1]$	RB
0	20	0.05	− 2.33	0.691	0.636	− 8.0	0.183	0.181	− 1.3
		0.25	− 0.95	0.666	0.622	− 6.5	0.173	0.172	− 0.3
		0.50	0.00	0.658	0.614	− 6.6	0.169	0.168	− 0.7
		0.75	0.95	0.667	0.623	− 6.7	0.173	0.172	− 0.3
		0.95	2.33	0.693	0.639	− 7.8	0.183	0.181	− 1.2
	50	0.05	− 2.33	0.595	0.539	− 9.5	0.174	0.174	− 0.1
		0.25	− 0.95	0.577	0.524	− 9.2	0.169	0.169	0.2
		0.50	0.00	0.572	0.520	− 9.2	0.168	0.168	0.3
		0.75	0.95	0.577	0.523	− 9.2	0.169	0.169	0.2
		0.95	2.33	0.595	0.537	− 9.7	0.174	0.174	− 0.1
3	20	0.05	− 1.17	0.096	0.156	62.4	0.170	0.081	− 52.5
		0.25	− 0.06	0.104	0.165	58.2	0.177	0.079	− 55.1
		0.50	0.73	0.111	0.173	55.3	0.187	0.079	− 58.0
		0.75	1.55	0.123	0.187	51.7	0.208	0.079	− 62.0
		0.95	2.76	0.147	0.206	40.3	0.276	0.081	− 70.8
	50	0.05	− 1.17	0.060	0.076	27.9	0.169	0.080	− 52.7
		0.25	− 0.06	0.066	0.082	23.4	0.176	0.079	− 55.0
		0.50	0.73	0.073	0.089	23.0	0.187	0.079	− 57.9
		0.75	1.55	0.081	0.098	21.4	0.208	0.079	− 62.0
		0.95	2.76	0.097	0.115	19.1	0.266	0.080	− 69.9

$$\mathrm{CMSE}_1 = g_{11}(\boldsymbol{\omega}, \boldsymbol{y}_{1(q)}) + \frac{1}{R}\sum_{r=1}^{R}\left(\hat{\theta}_1^{EB(r)} - \hat{\theta}_1^{B}\right)^2,$$

where $\boldsymbol{y}_{1(q)} = (y_{11(q)}, \ldots, y_{1n(q)})^\top$ and $\hat{\theta}_1^{EB(r)}$ is the EB estimate of θ_1 in the rth iteration. After preparing the true values of the CMSE, we calculate the mean of the estimates for the CMSE and the percentage relative bias (RB) based on $R = 1000$ simulation runs with each 100 bootstrap samples. These values are respectively calculated as

$$E\big[\widehat{\mathrm{CMSE}}_1\big] = \frac{1}{R}\sum_{r=1}^{R}\widehat{\mathrm{CMSE}}_1^{(r)},$$

$$\mathrm{RB}\ (\%) = 100 \times \frac{E\big[\widehat{\mathrm{CMSE}}_1\big] - \mathrm{CMSE}_1}{\mathrm{CMSE}_1},$$

where $\widehat{\mathrm{CMSE}}_1^{(r)}$ is the estimate of the CMSE in the rth replication. Those values are reported in Table 2. For comparison, we calculate the corresponding quantities of the estimates for the CMSE and the percentage relative bias (RB) under the standard NER model where both the random effect and the error term are assumed to be normally distributed. The results of the proposed skew-normal model and the usual NER model are respectively reported in the columns 'SN' and 'Normal' in Table 2.

As for the true value of the CMSE in the SN column of Table 2, it can be seen that the CMSE in the case of $\lambda = 0$ takes larger values than that in the case of $\lambda = 3$. Also, the CMSE decreases insignificantly as m increases when $\lambda = 0$. These behaviors may be caused by considerable variation of the estimators in the case of $\boldsymbol{a} = 0$. The result of the RB implies that the CMSE tends to be overestimated in the case of $\lambda = 3$, though getting more samples suppress this inflation. Also, the effect of the conditioning value is not negligible when $\lambda = 3$: the RB decreases as the conditioning values are located to the right. Comparing the results of the two methods for $\lambda = 0$ case, the proposed model gives larger true CMSE and the performance of the CMSE estimate is worse, which is a reasonable consequence caused by the redundancy of the skewness parameter λ. For $\lambda = 3$ case, however, the true CMSE is smaller and the effect of the sample size on reduction of both true CMSE and RB is more significant in the skew-normal model. When we use the normal model in the case of $\lambda = 3$, in contrast, the true CMSE cannot be reduced by increasing the sample size, which may result from inconsistency of the estimators for the parameters. Moreover, the CMSE is seriously underestimated regardless of the sample size. These results motivate us to use the proposed model in the situations where the skewness is present.

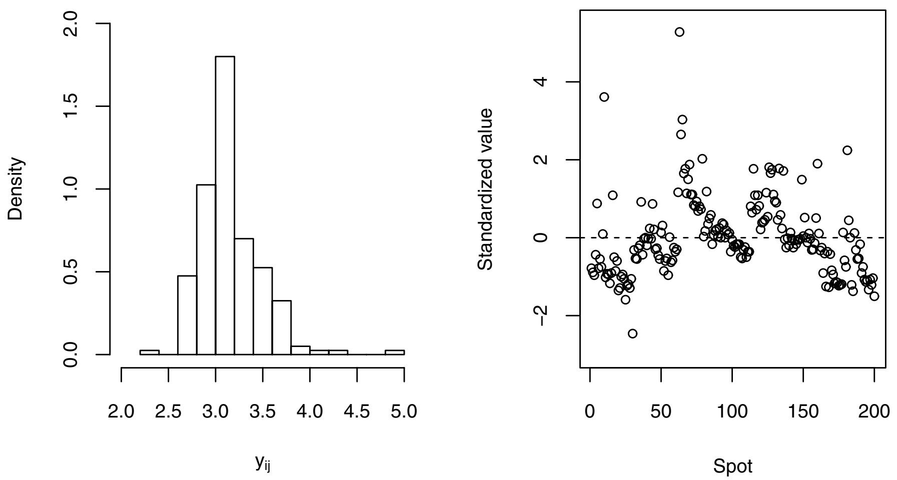

Fig. 2 Histogram of y_{ij} (left) and dot plot of the standardized value of y_{ij} (right) for the areas with $n_i \geq 3$

Table 3 The estimates of parameters and their standard errors

	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\sigma}^2$	$\hat{\tau}^2$	$\hat{\lambda}$
SN							
Estimate	5.281	– 0.631	– 0.068	0.139	0.036	0.008	1.936
s.e.	0.029	0.003	0.001	0.002	0.031	0.003	4.677
Normal							
Estimate	5.415	– 0.631	– 0.068	0.139	0.018	0.008	
s.e.	0.021	0.003	0.001	0.002	0.004	0.003	

6 An illustrative example

We investigate the performance of the suggested model and estimation methods through an illustrative example. We use the posted land price data along the Keikyu train line in 2001. This train line connects the suburbs in the Kanagawa prefecture and the metropolitan area in Tokyo, so that the commuters and students who live in the suburb area in the Kanagawa prefecture take this line to go to Tokyo. Thus the land price is expected to depend on the distance from Tokyo. The posted land price data are available for 52 stations on the Keikyu line. We consider each station as a small area, namely $m = 52$, and for the ith station, the land price data of n_i spots are available. Of the 52 stations, 37 stations have at least 3 observations.

For $j = 1, \ldots, n_i$, we have a set of observations $(y_{ij}, T_i^*, D_{ij}^*, FAR_{ij}^*)$, where y_{ij} denotes the log-transformed value of the posted land price (Yen/10,000) per square meter of the jth spot, T_i^* is the time (minute) it takes from the nearby station i to Tokyo station around 8:30 by train, D_{ij}^* is the geographical distance (meter) between the spot j and the station i and FAR_{ij}^* denotes the floor-area-ratio (%) of the spot j. Let

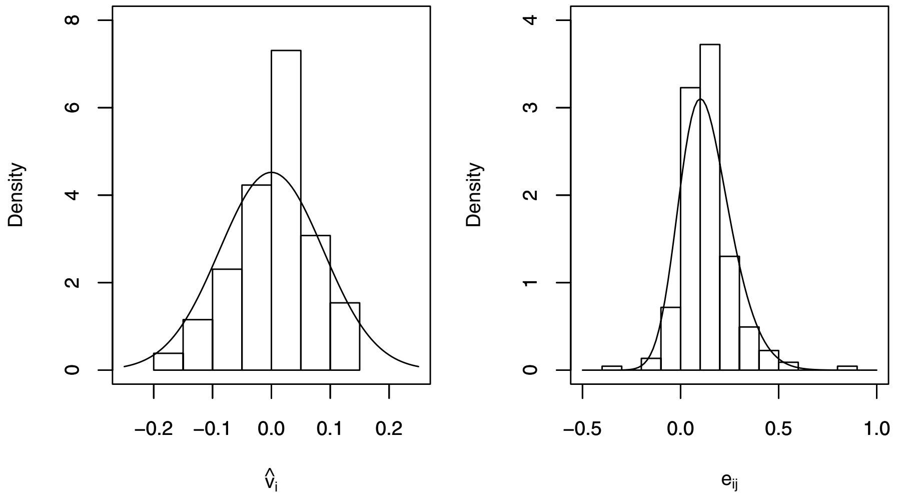

Fig. 3 Histogram of $\hat{v}_i$ (left) and e_{ij} (right) with the density of $\mathcal{N}(0, \hat{\tau}^2)$ (left) and $\mathcal{SN}(0, \hat{\sigma}^2, \widehat{\lambda})$ (right) superimposed

T_i, D_{ij} and FAR_{ij} be the log-transformed values of T_i^*, D_{ij}^* and FAR_{ij}^*. Then we consider the NER model with skew-normal errors, described as

$$y_{ij} = \beta_0 + T_i\beta_1 + D_{ij}\beta_2 + FAR_{ij}\beta_3 + v_i + \varepsilon_{ij},$$

where v_i's and ε_{ij}'s are mutually independent and distributed as $\mathcal{N}(0, \tau^2)$ and $\mathcal{SN}(0, \sigma^2, \lambda)$, respectively.

Table 4 Predicted values of θ_i and the CMSE using the proposed model (SN) and usual NER model (Normal) for the selected 12 areas

Area	n_i	SN		Normal	
		EB	$\widehat{\text{CMSE}}_1$	EB	$\widehat{\text{CMSE}}_1$
25	1	3.28	0.816	3.41	0.636
32	1	3.15	0.776	3.28	0.615
2	2	2.72	0.639	2.85	0.466
24	3	2.89	0.554	3.01	0.372
49	6	2.73	1.292	2.90	0.240
43	7	2.75	0.924	2.91	0.213
29	7	3.15	0.440	3.27	0.221
5	7	2.77	0.396	2.90	0.212
18	8	3.04	0.394	3.16	0.189
52	10	2.67	0.527	2.79	0.159
41	11	3.04	0.340	3.17	0.147
10	18	2.95	0.435	3.06	0.103

Values of $\widehat{\text{CMSE}}_1$ are multiplied by 100

The left panel of Fig. 2 gives the histogram of y_{ij} for those areas satisfying $n_i \geq 3$. It is revealed that the histogram of y_{ij} has a right-skewed shape, which means that after taking logarithm of the land price there still exists positive skewness. Hence, in this case, it is not appropriate to model the data based on symmetric distributions. However, the sample skewness of y_{ij}'s for the areas with $n_i \geq 3$ is 1.317, which is out of the feasible range of the population skewness described in Section 2.1. The plot of the standardized values of y_{ij} for $n_i \geq 3$ areas is depicted in Fig. 2, and it is observed that a couple of observations take extremely large values, which yields the strong positive-skewness. In our analysis given below, we remove the first and second largest observations from our analysis, and the resulting sample skewness drops to $\widetilde{\delta} = 0.637$.

The estimated values of the model parameters and their standard errors under the proposed model (SN) and the usual NER (Normal) are provided in Table 3, where the standard errors are provided by square roots of the jackknife variance estimators. The values for the two models are identical except for β_0, σ^2, and λ. The effect of T_i and D_{ij} are negative, which implies that the land price decreases as the time for going to Tokyo and/or the distance to the nearby station increase. The estimate of λ means that the land price data is still skewed to the right after log-transformation, though its standard error is so large. Since in the estimation of λ, we use the data of 37 areas with $n_i \geq 3$, we need more data to lower the standard error of $\widehat{\lambda}$. The histogram of the predicted random effect and the residual of the proposed model , denoted as $\hat{v}_i = E[v_i \mid \boldsymbol{y}_i]|_{\boldsymbol{\omega}=\hat{\boldsymbol{\omega}}}$ and $e_{ij} = y_{ij} - \boldsymbol{x}_{ij}^\top \widehat{\boldsymbol{\beta}} - \hat{v}_i$ respectively, are presented in Fig. 3. Using the estimates of parameters, the density of $\mathcal{N}(0, \hat{\tau}^2)$ for $\hat{v}_i$ and $\mathcal{SN}(0, \hat{\sigma}^2, \widehat{\lambda})$ for e_{ij} are depicted in each panel in Fig. 3. The sample coefficient of skewness is -0.33 for $\hat{v}_i$ and 1.02 for e_{ij}. Though $\hat{v}_i$ has a week negative skewness, departure of its histogram from the normal density does not seem to be so great. Thus it is reasonable to assume the normal distribution on the random effect. For e_{ij}, positive skewness is observed both in its skewness value and histogram in Fig. 3.

Next we predict (estimate) the land price of a spot with specific values of covariates. As covariates, we set a distance from the nearby station as 1000 m and a floor-area-ratio as 100%. Thus for each i, we want to predict

$$\theta_i = \beta_0 + T_i\beta_1 + D_0\beta_2 + FAR_0\beta_3 + v_i,$$

for $D_0 = \log(1000)$ and $FAR_0 = \log(100)$. The predicted values of θ_i and their CMSE estimates for 12 selected areas based on 2000 bootstrap samples are provided as the column 'SN' in Table 4. These quantities are also calculated under the standard NER model and reported in Table 4 as the column 'Normal'. It is easily observed that the EB estimates under the proposed model are smaller than those obtained by the usual NER model. As described in Sect.s 2.2 and 4.1, the major source of this gap is the difference between the estimates of $\widehat{\beta}_0$ for the two models, which equals to 0.134. Comparing the estimates of the CMSE, the skew-normal model gives larger values and, especially, in some areas the estimated CMSE in our model are several times larger than those in the normal model. Since the given data has strong positive skewness, it is considered that the values in the 'SN' and 'Normal' columns are respectively overestimated and underestimated. This point is illustrated in the simulation study presented in the previous section. Concerning the effect of area sample size, from the decreasing value of the CMSE estimate in the ‘Normal’ column it is suggested that the CMSE obtained under the normality assumption is a decreasing function of n_i. In general, decreasing tendency like this is not observed for the conditional MSE but for the unconditional MSE because the former usually also depends on $\boldsymbol{y}_i$. As shown by Booth and Hobert (1998), however, the leading term of the CMSE does not depend on $\boldsymbol{y}_i$ only in the model with normality assumption, so that increasing area sample size leads to reduction of the CMSE estimates in this case. In contrast, the estimated values of the CMSE for the proposed method do not appear to depend solely on n_i.

7 Concluding remarks

In this paper, we have considered the NER model with skew-normally distributed errors. Under this model, the Bayesian calculation has been conducted to provide simple forms of the posterior mean and variance. We have constructed the estimators of the model parameters similar to the method by Fuller and Battese (1973) and obtained the EB estimator. The uncertainty of the EB estimator has been studied in terms of the CMSE and the second-order unbiased estimator of the CMSE has been derived with the parametric bootstrap method. In the simulation study, it has been revealed that the proposed method yields the EB estimates with greater accuracy than the standard NER model does when the underlying distribution of the error term has skewness.

These results suggest that the proposed skew-normal model is useful if skewness is observed in the given data. However, it is a major problem that a skew-normal distribution cannot apply to highly skewed situations, which actually happened in our example in Section 6. This issue may be overcome by adopting more heavy-tailed distributions like a skew-*t* distribution and remains to be solved in the future works.

Acknowledgements We would like to thank the Associate Editor and the two reviewers for many valuable comments and helpful suggestions which led to an improved version of this paper. The research of the second author was supported in part by Grant-in-Aid for Scientific Research from Japan Society for the Promotion of Science (# 18K11188, # 15H01943 and # 26330036).

Appendix: Proofs

All the proofs of lemmas and theorems given in the paper are provided here.

Proof of expression (6) We explain briefly the derivation of the expression (6). The exponent in $f(v_i, \boldsymbol{u}_{1i} \mid \boldsymbol{y}_i)$ is proportional to

$$\frac{1+\lambda^2}{\sigma^2}\sum_{j=1}^{n_i}\left\{v_i - \left(y_{ij} - \boldsymbol{x}_{ij}^\top\boldsymbol{\beta} - \frac{\sigma\lambda}{\sqrt{1+\lambda_i}}u_{1ij}\right)\right\}^2 + \frac{v_i^2}{\tau^2} + \sum_{j=1}^{n_i}u_{1ij}^2,$$

which is rewritten as

$$\frac{(v_i-\mu_{v_i})^2}{\sigma_{v_i}^2}+\frac{1+\lambda^2}{\sigma^2}\left[\sum_{j=1}^{n_i}\left\{\frac{\sigma\lambda}{\sqrt{1+\lambda^2}}u_{1ij}-(y_{ij}-\boldsymbol{x}_{ij}^\top\boldsymbol{\beta})\right\}^2+\frac{\sigma^2}{1+\lambda^2}\sum_{j=1}^{n_i}u_{1ij}^2\right.$$
$$\left.-\frac{(1+\lambda^2)\tau^2}{\sigma^2+(1+\lambda^2)\tau^2 n_i}\left\{\frac{\sigma\lambda}{\sqrt{1+\lambda^2}}\sum_{j=1}^{n_i}u_{1ij}-\sum_{j=1}^{n_i}(y_{ij}-\boldsymbol{x}_{ij}^\top\boldsymbol{\beta})\right\}^2\right].$$

The first part corresponds to the density $\phi(v_i;\mu_{v_i},\sigma_{v_i}^2)$. For simplicity, let $\boldsymbol{J}_{n_i}=\boldsymbol{1}_{n_i}\boldsymbol{1}_{n_i}^\top$,

$$\boldsymbol{c}_i=\frac{\sigma\lambda}{\sqrt{1+\lambda^2}}(\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta})-\frac{n_i\sigma\lambda}{\sqrt{1+\lambda^2}}(\bar{y}_i-\bar{\boldsymbol{x}}_i^\top\boldsymbol{\beta})\boldsymbol{1}_{n_i}\quad\text{and}\quad d=\frac{\tau^2\sigma^2\lambda^2}{\sigma^2+n_i(1+\lambda^2)\tau^2}.$$

Then the second term can be expressed as $\sigma^2\boldsymbol{u}_{1i}^\top\boldsymbol{u}_{1i}-2\boldsymbol{c}_i^\top\boldsymbol{u}_{1i}-d\boldsymbol{u}_{1i}^\top\boldsymbol{J}_{n_i}\boldsymbol{u}_{1i}$. After completing square, one gets $\{\boldsymbol{u}_{1i}-(\sigma^2\boldsymbol{I}_{n_i}-d\boldsymbol{J}_{n_i})^{-1}\boldsymbol{c}_i\}^\top(\sigma^2\boldsymbol{I}_{n_i}-d\boldsymbol{J}_{n_i})\{\boldsymbol{u}_{1i}-(\sigma^2\boldsymbol{I}_{n_i}-d\boldsymbol{J}_{n_i})^{-1}\boldsymbol{c}_i\}$, which corresponds to the density $\phi_{n_i}(\boldsymbol{u}_{1i};\boldsymbol{\mu}_i,\sigma_{u_i}^2\boldsymbol{R}_i)$. Thus, we have the expression given in (6).

□

Proof of Lemma 3.1 Following Tallis (1961), we have

$$E[w_{ij}\,|\,\boldsymbol{y}_i]=\sum_{k=1}^{n_i}\rho_{i,jk}\phi(a_{ik})\frac{\int_{-\boldsymbol{a}_{i(-k)}}^{\infty}\phi_{n_i-1}(\boldsymbol{w}_i^k;\boldsymbol{0},\boldsymbol{R}_i^k)\,\mathrm{d}\boldsymbol{w}_{i(-k)}}{\int_{-\boldsymbol{a}_i}^{\infty}\phi_{n_i}(\boldsymbol{w}_i;\boldsymbol{0},\boldsymbol{R}_i)\,\mathrm{d}\boldsymbol{w}_i},\tag{19}$$

where $\boldsymbol{a}_{i(-k)}$ and $\boldsymbol{w}_{i(-k)}$ are respectively (n_i-1)-dimensional vector obtained by dropping the kth element of $\boldsymbol{a}_i$ and $\boldsymbol{w}_i$, $\boldsymbol{w}_i^k=(\boldsymbol{w}_{i(-k)}+\rho_i a_{ik}\boldsymbol{1}_{n_i-1})(1-\rho_i^2)^{-1/2}$, and $\boldsymbol{R}_i^k$ is the matrix of the partial correlation coefficients for $\boldsymbol{w}_i$. Using results from Dunnett and Sobel (1955), we reduce the two multiple integrals in (19) to one-dimensional integrals. Since the denominator of the fraction in (19) is written as

$$\int_{-\boldsymbol{a}_i}^{\infty}\phi_{n_i}(\boldsymbol{w}_i;\boldsymbol{0},\boldsymbol{R}_i)\,\mathrm{d}\boldsymbol{w}_i=\Pr\{W_j>-a_{ij},\,j=1,\dots,n_i\},$$

where $\boldsymbol{W}=(W_1,\dots,W_{n_i})^\top\sim\mathcal{N}_{n_i}(\boldsymbol{0},\boldsymbol{R}_i)$ with $(\boldsymbol{R}_i)_{qr}=\rho_i\in[0,1)$ for $q\neq r$. Thus W_j can be represented as

$$W_j=\sqrt{\rho_i}\xi_0+\sqrt{1-\rho_i}\xi_j,\quad j=1,\dots,n_i,\tag{20}$$

where for $j = 0, 1, \dots, n_i$, ξ_j's are mutually independently distributed as $\mathcal{N}(0, 1)$. This transformation gives

$$\Pr\{W_j > -a_{ij}, j = 1, \dots, n_i\} = \Pr\left\{\xi_j > \frac{-a_{ij} - \sqrt{\rho_i}\xi_0}{\sqrt{1-\rho_i}}, j = 1, \dots, n_i\right\}$$

$$= \int_{-\infty}^{\infty}\left\{\prod_{j=1}^{n_i}\left(1 - \Phi\left(\frac{-a_{ij} - \sqrt{\rho_i}\xi_0}{\sqrt{1-\rho_i}}\right)\right)\right\}\phi(\xi_0)\,\mathrm{d}\xi_0 = \int_{-\infty}^{\infty}\left\{\prod_{j=1}^{n_i}\Phi\left(\frac{a_{ij} + \sqrt{\rho_i}\xi_0}{\sqrt{1-\rho_i}}\right)\right\}\phi(\xi_0)\,\mathrm{d}\xi_0,$$

which corresponds to (8).

Similarly, we see that the numerator in (19) is

$$\int_{-\boldsymbol{a}_{i(-k)}}^{\infty}\phi_{n_i-1}(\boldsymbol{w}_i^k;\boldsymbol{0},\boldsymbol{R}_i^k)\,\mathrm{d}\boldsymbol{w}_{i(-k)} = \Pr\left\{W_l^k > \frac{-a_{il} + \rho_i a_{ik}}{(1-\rho_i^2)^{1/2}},\ l = 1, \dots, n_i,\ l \neq k\right\},$$

where $\boldsymbol{W}^k = (W_1^k, \dots, W_{k-1}^k, W_{k+1}^k, \dots, W_{n_i}^k)^\top \sim \mathcal{N}_{n_i-1}(\boldsymbol{0}, \boldsymbol{R}_i^k)$ with $(\boldsymbol{R}_i^k)_{qr} = \rho_i/(1+\rho_i)$ for $q \neq r$. Thus, analogously to (20), W_l^k can be expressed as

$$W_l^k = \sqrt{\frac{\rho_i}{1+\rho_i}}\xi_0 + \sqrt{\frac{1}{1+\rho_i}}\xi_l, \quad l = 1, \dots, n_i,\ l \neq k,$$

where ξ_l's are mutually independent standard normal variables again. Then it follows that

$$\Pr\left\{W_l^k > \frac{-a_{il} + \rho_i a_{ik}}{(1-\rho_i^2)^{1/2}},\ l = 1, \dots, n_i,\ l \neq k\right\}$$

$$= \int_{-\infty}^{\infty}\left\{\prod_{l\neq k}\left(1 - \Phi\left(\frac{-a_{il} + \rho_i a_{ik}}{\sqrt{1-\rho_i}} - \sqrt{\rho_i}\xi_0\right)\right)\right\}\phi(\xi_0)\,\mathrm{d}\xi_0$$

$$= \int_{-\infty}^{\infty}\left\{\prod_{l\neq k}\Phi\left(\frac{a_{il} - \rho_i a_{ik}}{\sqrt{1-\rho_i}} + \sqrt{\rho_i}\xi_0\right)\right\}\phi(\xi_0)\,\mathrm{d}\xi_0,$$

which gives (9). □

Proof of Theorem 3.1 It suffices to transform the second term of (7) into the desired form. Using Lemma 3.1 and the fact that $(\boldsymbol{R}_i)_{jk} = \rho_i$ for $j \neq k$, we have

$$\begin{aligned}&\frac{n_i\tau^2(1+\lambda^2)}{\sigma^2 + n_i\tau^2(1+\lambda^2)}\sigma_{u_i}\sum_{j=1}^{n_i}E[w_{ij}\,|\,\boldsymbol{y}_i]\\&\quad= \frac{n_i\tau^2(1+\lambda^2)}{\sigma^2 + n_i\tau^2(1+\lambda^2)}\sigma_{u_i}(1 + (n_i - 1)\rho_i)\sum_{j=1}^{n_i}\phi(a_{ij})\frac{\alpha_{1ij}}{\alpha_{0i}}.\end{aligned} \tag{21}$$

Since

$$1+(n_i-1)\rho_i = \frac{1}{(1+\lambda^2)\sigma_{u_i}^2}\frac{\sigma^2+n_i\tau^2(1+\lambda^2)}{\sigma^2+n_i\tau^2}, \tag{22}$$

the formula (21) reduces to

$$\frac{n_i\tau^2}{\sigma^2+n_i\tau^2}\sum_{j=1}^{n_i}\sigma_{u_i}^{-1}\phi(a_{ij})\frac{\alpha_{1ij}}{\alpha_{0i}},$$

which gives the desired expression. □

Proof of Lemma 3.2 The outline is the same as in the proof of Lemma 3.1. Using the results of Tallis (1961) and Theorem 3.1, we have

$$\begin{aligned} E[w_{ij}w_{ik}\,|\,\boldsymbol{y}_i] = \rho_{i,jk} &- \sum_{q=1}^{n_i}\rho_{i,jq}\rho_{i,kq}a_{iq}\phi(a_{iq})\frac{\alpha_{1iq}}{\alpha_{0i}} \\ &+ \sum_{q=1}^{n_i}\rho_{i,jq}\sum_{r\neq q}(\rho_{i,kr}-\rho_i\rho_{i,kq})\phi(a_{iq},a_{ir};\rho_i)\alpha_{0i}^{-1} \\ &\times\int_{-\boldsymbol{a}_{i(-q,r)}}^{\infty}\phi_{n_i-2}(\boldsymbol{w}_i^{qr};\boldsymbol{0},\boldsymbol{R}_i^{qr})\,\mathrm{d}\boldsymbol{w}_{i(-q,r)}, \end{aligned} \tag{23}$$

where $\boldsymbol{a}_{i(-q,r)}$ and $\boldsymbol{w}_{i(-q,r)}$ are respectively (n_i-2)-dimensional vectors obtained by dropping the qth and rth elements of $\boldsymbol{a}$ and $\boldsymbol{w}$,

$$\boldsymbol{w}_i^{qr} = \frac{\boldsymbol{w}_{i(-q,r)}+\rho_i(1+\rho_i)^{-1}(a_{iq}+a_{ir})\mathbf{1}_{n_i-2}}{\sqrt{(1-\rho_i)(1+2\rho_i)(1+\rho_i)^{-1}}},$$

and $\boldsymbol{R}_i^{qr}$ is the matrix of the second-order partial correlation coefficients for $\boldsymbol{w}_i$. The (n_i-2)-fold integral in (23) is written as

$$\begin{aligned} &\int_{-\boldsymbol{a}_{i(-q,r)}}^{\infty}\phi_{n_i-2}(\boldsymbol{w}_i^{qr};\boldsymbol{0},\boldsymbol{R}_i^{qr})\,\mathrm{d}\boldsymbol{w}_{i(-q,r)} \\ &= \Pr\left\{W_s^{qr} > \frac{-a_{is}+\rho_i(1+\rho_i)^{-1}(a_{iq}+a_{ir})}{\sqrt{(1-\rho_i)(1+2\rho_i)(1+\rho_i)^{-1}}},\ s=1,\dots,n_i,\ s\neq q,r\right\} \end{aligned}$$

where $\boldsymbol{W}^{qr} \sim \mathcal{N}_{n_i-2}(\mathbf{0}, \boldsymbol{R}_i^{qr})$ with $(\boldsymbol{R}_i^{qr})_{st} = \rho_i/(1+2\rho_i)$ for $s \neq t$. Here the definition of $\boldsymbol{W}^{qr}$ is analogous to $\boldsymbol{W}^{q}$ in the proof of Lemma 3.1. Then using the similar method to (20) with $\rho_i/(1+2\rho_i)$ instead of ρ_i, we have

$$W_s^{qr} = \sqrt{\frac{\rho_i}{1+2\rho_i}}\xi_0 + \sqrt{\frac{1+\rho_i}{1+2\rho_i}}\xi_s, \quad s = 1, \dots, n_i,\ s \neq q, r$$

where ξ_s's are mutually independent standard normal random variables. Then we have

$$\begin{aligned}
&\Pr\left\{ W_s^{qr} > \frac{-a_{is} + \rho_i(1+\rho_i)^{-1}(a_{iq}+a_{ir})}{\sqrt{(1-\rho_i)(1+2\rho_i)(1+\rho_i)^{-1}}},\ s = 1, \dots, n_i,\ s \neq q, r \right\} \\
&= \int_{-\infty}^{\infty} \left\{ \prod_{s \neq q,r} \left(1 - \Phi\left(\frac{-a_{is} + \rho_i(1+\rho_i)^{-1}(a_{iq}+a_{ir})}{\sqrt{1-\rho_i}} - \sqrt{\frac{\rho_i}{1+\rho_i}}\xi_0 \right)\right)\right\} \phi(\xi_0)\,\mathrm{d}\xi_0 \\
&= \int_{-\infty}^{\infty} \left\{ \prod_{s \neq q,r} \Phi\left(\frac{a_{is} - \rho_i(1+\rho_i)^{-1}(a_{iq}+a_{ir})}{\sqrt{1-\rho_i}} + \sqrt{\frac{\rho_i}{1+\rho_i}}\xi_0 \right)\right\} \phi(\xi_0)\,\mathrm{d}\xi_0,
\end{aligned}$$

which corresponds to (12). □

Proof of Theorem 3.2 It follows from Lemmas 3.1 and 3.2 that

$$\begin{aligned}
\sum_{j=1}^{n_i} E[w_{ij} \mid \boldsymbol{y}_i] &= (1 + (n_i - 1)\rho_i) \sum_{q=1}^{n_i} \phi(a_{iq}) \frac{\alpha_{1iq}}{\alpha_{0i}}, \\
\sum_{j=1}^{n_i} E[w_{ij}^2 \mid \boldsymbol{y}_i] &= n_i - (1 + (n_i - 1)\rho_i)^2 \sum_{q=1}^{n_i} a_{iq}\phi(a_{iq}) \frac{\alpha_{1iq}}{\alpha_{0i}} \\
&\quad + \rho_i(1-\rho_i)(1 + (n_i - 1)\rho_i) \sum_{q=1}^{n_i} \sum_{r \neq q} \phi(a_{iq}, a_{ir}; \rho_i) \frac{\alpha_{2iqr}}{\alpha_{0i}}, \\
\sum_{j=1}^{n_i} \sum_{k \neq j} E[w_{ij}w_{ik} \mid \boldsymbol{y}_i] &= n_i(n_i-1)\rho_i - (n_i - 1)\rho_i(2 + (n_i - 2)\rho_i) \sum_{q=1}^{n_i} a_{iq}\phi(a_{iq}) \frac{\alpha_{1iq}}{\alpha_{0i}} \\
&\quad + (1-\rho_i)(1 + (n_i - 1)\rho_i)(1 + (n_i - 2)\rho_i) \sum_{q=1}^{n_i} \sum_{r \neq q} \phi(a_{iq}, a_{ir}; \rho_i) \frac{\alpha_{2iqr}}{\alpha_{0i}}.
\end{aligned}$$

The conditional variance of $\sum_{j=1}^{n_i} w_{ij}$ given $\boldsymbol{y}_i$ is

$$\begin{aligned}\mathrm{Var}\Big(\sum_{j=1}^{n_i} w_{ij}\,\Big|\,\boldsymbol{y}_i\Big) &= \sum_{j=1}^{n_i} E[w_{ij}^2\,|\,\boldsymbol{y}_i] + \sum_{j=1}^{n_i}\sum_{k\neq j} E[w_{ij}w_{ik}\,|\,\boldsymbol{y}_i] - \Big\{\sum_{j=1}^{n_i} E[w_{ij}\,|\,\boldsymbol{y}_i]\Big\}^2 \\ &= n_i(1+(n_i-1)\rho_i) - (1+(n_i-1)\rho_i)^2 \nu_i(\boldsymbol{\omega},\boldsymbol{y}_i),\end{aligned} \tag{24}$$

where $\nu_i(\boldsymbol{\omega},\boldsymbol{y}_i)$ is defined as (14). Then it follows from (11), (22) and (24) that

$$\begin{aligned}\mathrm{Var}(\theta_i\,|\,\boldsymbol{y}_i) = &\frac{\sigma^2\tau^2}{\sigma^2+n_i\tau^2(1+\lambda^2)} + \frac{n_i\sigma^2\tau^4\lambda^2}{(\sigma^2+n_i\tau^2(1+\lambda^2))(\sigma^2+n_i\tau^2)} \\ &- \frac{1}{(1+\lambda^2)\sigma_{u_i}^2}\frac{\sigma^2\tau^4\lambda^2}{(\sigma^2+n_i\tau^2)^2}\nu_i(\boldsymbol{\omega},\boldsymbol{y}_i) \\ = &\frac{\sigma^2\tau^2}{\sigma^2+n_i\tau^2(1+\lambda^2)}\Big(1+\frac{n_i\tau^2\lambda^2}{\sigma^2+n_i\tau^2}\Big) \\ &- \frac{1}{1+\tau^2\lambda^2/(\sigma^2+n_i\tau^2)}\frac{\sigma^2\tau^4\lambda^2}{(\sigma^2+n_i\tau^2)^2}\nu_i(\boldsymbol{\omega},\boldsymbol{y}_i) \\ = &\frac{\sigma^2\tau^2}{\sigma^2+n_i\tau^2} - \frac{\sigma^2\tau^2}{\sigma^2+n_i\tau^2}\frac{\tau^2\lambda^2/(\sigma^2+n_i\tau^2)}{1+\tau^2\lambda^2/(\sigma^2+n_i\tau^2)}\nu_i(\boldsymbol{\omega},\boldsymbol{y}_i) \\ = &\frac{\sigma^2\tau^2}{\sigma^2+n_i\tau^2}\{1-\rho_i\nu_i(\boldsymbol{\omega},\boldsymbol{y}_i)\},\end{aligned}$$

which proves Theorem 3.2. □

Proof of Theorem 4.1 First, we derive the desired properties for the parameters $\big(\boldsymbol{\beta}_\varepsilon^\top, \sigma^2, \tau^2, \lambda\big)$. In general, consider the case that two estimators $\hat{\theta}_1$ of θ_1 and $\hat{\theta}_2$ of θ_2 have the forms

$$\begin{aligned}\hat{\theta}_1 - \theta_1 &= \frac{1}{m}\sum_{i=1}^m h_{1i}(\boldsymbol{y}_i), \\ \hat{\theta}_2 - \theta_2 &= \frac{1}{m}\sum_{i=1}^m h_{2i}(\boldsymbol{y}_i),\end{aligned}$$

where $h_{1i}(\boldsymbol{y}_i)$ and $h_{2i}(\boldsymbol{y}_i)$ (written as h_{1i} and h_{2i} for simplicity) are functions of $\boldsymbol{y}_i$ such that $h_{ki} = O_p(1)$, $E[h_{ki}] = O(1)$ for $k = 1, 2$. Since $\boldsymbol{y}_i$'s are mutually independent, it is shown that

$$\begin{aligned}
E[\hat{\theta}_1-\theta_1\,|\,\boldsymbol{y}_i] &= E[\hat{\theta}_1-\theta_1]+\frac{1}{m}(h_{1i}-E[h_{1i}]) = E[\hat{\theta}_1-\theta_1]+O_p(m^{-1}),\\
E[(\hat{\theta}_1-\theta_1)(\hat{\theta}_2-\theta_2)\,|\,\boldsymbol{y}_i] &= E[(\hat{\theta}_1-\theta_1)(\hat{\theta}_2-\theta_2)]\\
&\quad+\frac{1}{m}\big\{(h_{1i}-E[h_{1i}])E[\hat{\theta}_2-\theta_2]+(h_{2i}-E[h_{2i}])E[\hat{\theta}_1-\theta_1]\big\}\\
&\quad+\frac{1}{m^2}\big\{(h_{1i}-E[h_{1i}])(h_{2i}-E[h_{2i}])-E[(h_{1i}-E[h_{1i}])(h_{2i}-E[h_{2i}])]\big\}\\
&= E[(\hat{\theta}_1-\theta_1)(\hat{\theta}_2-\theta_2)]+O_p(m^{-1}),
\end{aligned}$$

which means that it is enough to obtain the required results for the unconditional expectations. Note that all the moments of a skew-normal distribution exist. Since the methods of estimating $\boldsymbol{\beta}_\varepsilon$ and m_2 are the same as those of the usual NER model, it follows from the results of Fuller and Battese (1973) that

$$\begin{aligned}
E\big[\big(\widehat{\boldsymbol{\beta}}_\varepsilon-\boldsymbol{\beta}_\varepsilon\big)\big(\widehat{\boldsymbol{\beta}}_\varepsilon-\boldsymbol{\beta}_\varepsilon\big)^\top\big] &= O(m^{-1}), \quad E[(\widehat{m}_2-m_2)^2]=O(m^{-1}),\\
E\big[\widehat{\boldsymbol{\beta}}_\varepsilon-\boldsymbol{\beta}_\varepsilon\big] &= O(m^{-1}), \qquad E[\widehat{m}_2-m_2]=0.
\end{aligned}$$

The last formula comes from the unbiasedness of $\widehat{m}_2$.

Next we treat $\widehat{m}_3$. Let $\widehat{\boldsymbol{\beta}}_2^{FE}$ be the OLS estimator obtained by regressing y_{ij} on $\tilde{z}_{ij}$. Then, $\widehat{\boldsymbol{\beta}}_1^{FE}$ is written as $\widehat{\boldsymbol{\beta}}_2^{FE}=\boldsymbol{\beta}_2+(\sum_{i=1}^m\sum_{j=1}^{n_i}\tilde{z}_{ij}\tilde{z}_{ij}^\top)^{-1}\sum_{i=1}^m\sum_{j=1}^{n_i}\tilde{z}_{ij}\tilde{\varepsilon}_{ij}$. It follows from (RC) and $\widehat{\boldsymbol{\beta}}_2^{FE}-\boldsymbol{\beta}_2=O_p(m^{-1/2})$ that

$$\begin{aligned}
\widehat{m}_3 &= \frac{1}{\eta_1}\sum_{i=1}^m\sum_{j=1}^{n_i}\tilde{\varepsilon}_{ij}^3-\frac{3}{\eta_1}(\widehat{\boldsymbol{\beta}}_2^{FE}-\boldsymbol{\beta}_2)^\top\left(\sum_{i=1}^m\sum_{j=1}^{n_i}\tilde{z}_{ij}\tilde{\varepsilon}_{ij}^2\right)\\
&\quad+\frac{3}{\eta_1}(\widehat{\boldsymbol{\beta}}_2^{FE}-\boldsymbol{\beta}_2)^\top\left(\sum_{i=1}^m\sum_{j=1}^{n_i}\tilde{z}_{ij}\tilde{z}_{ij}^\top\tilde{\varepsilon}_{ij}\right)(\widehat{\boldsymbol{\beta}}_2^{FE}-\boldsymbol{\beta}_2)-\frac{1}{\eta_1}\sum_{i=1}^m\sum_{j=1}^{n_i}\{\tilde{z}_{ij}^\top(\widehat{\boldsymbol{\beta}}_2^{FE}-\boldsymbol{\beta}_2)\}^3\\
&= \frac{1}{\eta_1}\sum_{i=1}^m\sum_{j=1}^{n_i}\tilde{\varepsilon}_{ij}^3-\frac{3}{\eta_1}\left(\sum_{i=1}^m\sum_{j=1}^{n_i}\tilde{z}_{ij}^\top\tilde{\varepsilon}_{ij}\right)\left(\sum_{i=1}^m\sum_{j=1}^{n_i}\tilde{z}_{ij}\tilde{z}_{ij}^\top\right)^{-1}\left(\sum_{i=1}^m\sum_{j=1}^{n_i}\tilde{z}_{ij}\tilde{\varepsilon}_{ij}^2\right)+O_p(m^{-3/2}).
\end{aligned}$$

Since $\tilde{\varepsilon}_{ij}$'s are independent for different i and $E[\tilde{\varepsilon}_{ij}]=0$, the bias of $\widehat{m}_3$ is

$$\begin{aligned}
E[\widehat{m}_3-m_3] &= -\frac{3}{\eta_1}\sum_{i=1}^m E\Big[\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\tilde{z}_{ij}^\top\tilde{\varepsilon}_{ij}\Big(\sum_{q=1}^m\sum_{r=1}^{n_i}\tilde{z}_{qr}\tilde{z}_{qr}^\top\Big)^{-1}\tilde{z}_{ik}\tilde{\varepsilon}_{ik}^2\Big]+o(m^{-1})\\
&= -\frac{3}{m\eta_1}\sum_{i=1}^m\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\tilde{z}_{ij}^\top\Big(\frac{1}{m}\sum_{q=1}^m\sum_{r=1}^{n_i}\tilde{z}_{qr}\tilde{z}_{qr}^\top\Big)^{-1}\tilde{z}_{ik}E[\tilde{\varepsilon}_{ij}\tilde{\varepsilon}_{ik}^2]+o(m^{-1}),
\end{aligned}$$

which is of order $O(m^{-1})$. From the fact that $\widehat{m}_3 = \eta_1^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\tilde{\varepsilon}_{ij}^3 + O_p(m^{-1})$ and $\eta_1^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\tilde{\varepsilon}_{ij}^3 = O_p(m^{-1/2})$, it follows that

$$\begin{aligned} E[(\widehat{m}_3 - m_3)^2] &= E\left[\left(\frac{1}{\eta_1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\tilde{\varepsilon}_{ij}^3 - m_3\right)^2\right] + o(m^{-1}) \\ &= \frac{1}{\eta_1^2}\sum_{i=1}^{m} E\left[\left(\sum_{j=1}^{n_i}\tilde{\varepsilon}_{ij}^3 - \frac{(n_i-1)(n_i-2)}{n_i}m_3\right)^2\right] + o(m^{-1}). \end{aligned} \tag{25}$$

The expectation in (25) is bounded under the condition (RC) and the existence of up to the sixth moment of ε_{ij}, which leads to $E[(\widehat{m}_3 - m_3)^2] = O(m^{-1})$. Then we have $\widehat{m}_2 - m_2 = O_p(m^{-1/2})$ and $\widehat{m}_3 - m_3 = O_p(m^{-1/2})$. Also, $E[(\widehat{m}_2 - m_2)(\widehat{m}_3 - m_3)]$ can be treated by Schwarz's inequality as

$$|E[(\widehat{m}_2 - m_2)(\widehat{m}_3 - m_3)]| \le (E[(\widehat{m}_2 - m_2)^2]E[(\widehat{m}_3 - m_3)^2])^{1/2} = O(m^{-1}). \tag{26}$$

The inverse transformation of $(m_2(\sigma^2,\lambda), m_3(\sigma^2,\lambda))$ is derived from (1) and (2) as

$$\begin{aligned} \sigma^2(m_2,m_3) &= m_2 + \left(\frac{2}{4-\pi}m_3\right)^{2/3}, \\ \delta(m_2,m_3) &= \sqrt{\frac{\pi}{2}}\left(\frac{2}{4-\pi}m_3\right)^{1/3}\left\{m_2 + \left(\frac{2}{4-\pi}m_3\right)^{2/3}\right\}^{-1/2}. \end{aligned}$$

Since $\lambda \neq 0$, $\delta \neq 0$, or $m_3 \neq 0$, it is easy to check these functions are three times continuously differentiable. Thus, using the Taylor series expansion we have

$$\begin{aligned} \begin{pmatrix} \hat{\sigma}^2 - \sigma^2 \\ \tilde{\delta} - \delta \end{pmatrix} &= \begin{pmatrix} \partial\sigma^2/\partial m_2 & \partial\sigma^2/\partial m_3 \\ \partial\delta/\partial m_2 & \partial\delta/\partial m_3 \end{pmatrix} \begin{pmatrix} \widehat{m}_2 - m_2 \\ \widehat{m}_3 - m_3 \end{pmatrix} \\ &\quad + \frac{1}{2}\sum_{r=2}^{3}\left\{\frac{\partial}{\partial m_r}\begin{pmatrix} \partial\sigma^2/\partial m_2 & \partial\sigma^2/\partial m_3 \\ \partial\delta/\partial m_2 & \partial\delta/\partial m_3 \end{pmatrix}\right\} \begin{pmatrix} \widehat{m}_2 - m_2 \\ \widehat{m}_3 - m_3 \end{pmatrix} (\widehat{m}_r - m_r) + O_p(m^{-3/2}), \end{aligned}$$

which, together with the results obtained up to this point, gives

$$E\left[\begin{pmatrix} \hat{\sigma}^2 - \sigma^2 \\ \tilde{\delta} - \delta \end{pmatrix}\begin{pmatrix} \hat{\sigma}^2 - \sigma^2 \\ \tilde{\delta} - \delta \end{pmatrix}^{\top}\right] = O(m^{-1}), \quad E\left[\begin{pmatrix} \hat{\sigma}^2 - \sigma^2 \\ \tilde{\delta} - \delta \end{pmatrix}\right] = O(m^{-1}).$$

Concerning the truncated estimator $\widehat{\delta} = \max(-1 + 1/m, \min(\widetilde{\delta}, 1 - 1/m))$, we consider the case of $0 < \delta < 1$. For large m, we have $1 - 1/m - \delta > 0$. Then, $\Pr(\widetilde{\delta} > 1 - 1/m) = \Pr(\widetilde{\delta} - \delta > 1 - 1/m - \delta) \le E[(\widetilde{\delta} - \delta)^2]/(1 - 1/m - \delta)^2$, so that $\Pr(\widetilde{\delta} > 1 - 1/m) = O(m^{-1})$. This shows the consistency of $\widehat{\delta}$. Using the same arguments as below, we can show that $E[\widehat{\delta} - \delta] = O(m^{-1})$ and $E[(\widehat{\delta} - \delta)^2] = O(m^{-1})$. These results lead to the asymptotic properties of $\widehat{\lambda}$.

Concerning $\hat{\tau}^2$, we have $E[(\tilde{\tau}^2-\tau^2)^2]=O(m^{-1})$, because $(v_i,\,\varepsilon_{ij})$'s are independent for different i and all the moments of v_i and ε_{ij} exist. Following Prasad and Rao (1990),

$$\Pr\{\tilde{\tau}^2\le 0\}=\Pr\{\tilde{\tau}^2-\tau^2\le-\tau^2\}\le\Pr\{|\tilde{\tau}^2-\tau^2|\ge\tau^2\}\le\frac{E[(\tilde{\tau}^2-\tau^2)^2]}{\tau^4},\tag{27}$$

which is of order $O(m^{-1})$. Then we have

$$\begin{aligned}E[(\hat{\tau}^2-\tau^2)^2]&=E[(\tilde{\tau}^2-\tau^2)^2\mid\tilde{\tau}^2>0]\Pr\{\tilde{\tau}^2>0\}+E[(0-\tau^2)^2\mid\tilde{\tau}^2\le 0]\Pr\{\tilde{\tau}^2\le 0\}\\&\le E[(\tilde{\tau}^2-\tau^2)^2]+\tau^4\Pr\{\tilde{\tau}^2\le 0\},\end{aligned}$$

which is of order $O(m^{-1})$. Also, it follows from $E[\tilde{\tau}^2-\tau^2]=0$ that

$$\begin{aligned}E[\hat{\tau}^2-\tau^2]&=E[\tilde{\tau}^2I(\tilde{\tau}^2>0)-\tau^2]=E[\tilde{\tau}^2\{1-I(\tilde{\tau}^2\le 0)\}-\tau^2]\\&=E[-\tilde{\tau}^2I(\tilde{\tau}^2\le 0)]=E[(-\tilde{\tau}^2+\tau^2)I(-\tilde{\tau}^2+\tau^2\ge\tau^2)]-\tau^2\Pr\{\tilde{\tau}^2\le 0\}.\end{aligned}$$

As for the first term,

$$\begin{aligned}E[(-\tilde{\tau}^2+\tau^2)I(-\tilde{\tau}^2+\tau^2\ge\tau^2)]&=\tau^2E\Big[\frac{-\tilde{\tau}^2+\tau^2}{\tau^2}I\Big(\frac{-\tilde{\tau}^2+\tau^2}{\tau^2}\ge 1\Big)\Big]\\&\le\tau^2E\Big[\Big(\frac{-\tilde{\tau}^2+\tau^2}{\tau^2}\Big)^2I\Big(\frac{-\tilde{\tau}^2+\tau^2}{\tau^2}\ge 1\Big)\Big]\\&\le\frac{E[(\tilde{\tau}^2-\tau^2)^2]}{\tau^2},\end{aligned}$$

which is of order $O(m^{-1})$. Thus, together with (27), we obtain $E[\hat{\tau}^2-\tau^2]=O(m^{-1})$.

We have derived the desired properties for the unconditional case, so that from the argument at the beginning of the proof, the statement of the theorem has been checked for $\underline{\left(\boldsymbol{\beta}_\varepsilon^\top,\sigma^2,\tau^2,\right.}\!\left.\lambda\right)$. Lastly, we need to consider β_0 instead of $\beta_{0\varepsilon}$. Since $\mu_\varepsilon=\sigma\sqrt{2/\pi\lambda}/\sqrt{1+\lambda^2}$ is a three times continuously differentiable function of (σ^2,λ), it follows that

$$\begin{aligned}\widehat{\mu}_\varepsilon-\mu_\varepsilon&=\begin{pmatrix}\hat{\sigma}^2-\sigma^2\\ \widehat{\lambda}-\lambda\end{pmatrix}^\top\begin{pmatrix}\partial\mu_\varepsilon/\partial\sigma^2\\ \partial\mu_\varepsilon/\partial\lambda\end{pmatrix}\\&\quad+\frac{1}{2}\begin{pmatrix}\hat{\sigma}^2-\sigma^2\\ \widehat{\lambda}-\lambda\end{pmatrix}^\top\begin{pmatrix}\partial^2\mu_\varepsilon/(\partial\sigma^2\partial\sigma^2)&\partial^2\mu_\varepsilon/(\partial\lambda\partial\sigma^2)\\ \partial^2\mu_\varepsilon/(\partial\sigma^2\partial\lambda)&\partial^2\mu_\varepsilon/(\partial\lambda\partial\lambda)\end{pmatrix}\begin{pmatrix}\hat{\sigma}^2-\sigma^2\\ \widehat{\lambda}-\lambda\end{pmatrix}+O_p(m^{-3/2}).\end{aligned}$$

Using this expansion, the desired results can be easily obtained.

It remains to show the expectations of cross terms are $O(m^{-1})$. Analogously to (26), this can be achieved by Schwarz's inequality, and the proof is complete. □

Proof of Proposition 4.1 From Theorem 4.1, $g_{1i}(\widehat{\boldsymbol{\omega}},\boldsymbol{y}_i)$ can be expanded as $g_{1i}(\widehat{\boldsymbol{\omega}},\boldsymbol{y}_i)=g_{1i}(\boldsymbol{\omega},\boldsymbol{y}_i)+G_{1i}(\widehat{\boldsymbol{\omega}},\boldsymbol{\omega},\boldsymbol{y}_i)+O_p(m^{-3/2})$ where

$$G_{1i}(\widehat{\boldsymbol{\omega}}, \boldsymbol{\omega}, \boldsymbol{y}_i) = (\widehat{\boldsymbol{\omega}} - \boldsymbol{\omega})^\top \frac{\partial g_{1i}(\boldsymbol{\omega}, \boldsymbol{y}_i)}{\partial \boldsymbol{\omega}} + \frac{1}{2}(\widehat{\boldsymbol{\omega}} - \boldsymbol{\omega})^\top \frac{\partial^2 g_{1i}(\boldsymbol{\omega}, \boldsymbol{y}_i)}{\partial \boldsymbol{\omega} \partial \boldsymbol{\omega}^\top}(\widehat{\boldsymbol{\omega}} - \boldsymbol{\omega}).$$

Thus we have $E[g_{1i}(\widehat{\boldsymbol{\omega}}, \boldsymbol{y}_i) \,|\, \boldsymbol{y}_i] = g_{1i}(\boldsymbol{\omega}, \boldsymbol{y}_i) + E[G_{1i}(\widehat{\boldsymbol{\omega}}, \boldsymbol{\omega}, \boldsymbol{y}_i) \,|\, \boldsymbol{y}_i] + o_p(m^{-1})$. It follows from Theorem 4.1 that $E[G_{1i}(\widehat{\boldsymbol{\omega}}, \boldsymbol{\omega}, \boldsymbol{y}_i) \,|\, \boldsymbol{y}_i] = O_p(m^{-1})$, so that applying the same arguments as in Butar and Lahiri (2003) shows $E[\hat{g}_{1i} \,|\, \boldsymbol{y}_i] = g_{1i}(\boldsymbol{\omega}, \boldsymbol{y}_i) + o_p(m^{-1})$. Also, using Theorem 4.1 again, it can be seen that $E[\hat{g}_{i2} \,|\, \boldsymbol{y}_i] = g_{2i}(\boldsymbol{\omega}, \boldsymbol{y}_i) + o_p(m^{-1})$. Then the proposition can be immediately obtained. □

References

Arellano-Valle, R. B., Bolfarine, H., & Lachos, V. H. (2005). Skew-normal linear mixed models. *Journal of Data Science*, *3*, 415–438.

Arellano-Valle, R. B., Bolfarine, H., & Lachos, V. (2007). Bayesian inference for skew-normal linear mixed models. *Journal of Applied Statistics*, *34*, 663–682.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, *12*, 171–178.

Azzalini, A. (1986). Further results on a class of distributions which includes the normal ones. *Statistica*, *XLVI*, 199–208.

Azzalini, A. (2013). *The skew-normal and related families*. Cambridge: Cambridge University Press.

Azzalini, A., & Capitanio, A. (1999). Statistical applications of the multivariate skew normal distribution. *Journal of the Royal Statistical Society*, *B 61*, 579–602.

Battese, G. E., Harter, R. M., & Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, *83*, 28–36.

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

Butar, F. B., & Lahiri, P. (2003). On measures of uncertainty of empirical Bayes small-area estimators. *Journal of Statistical Planning and Inference*, *112*, 63–76.

Diallo, M., & Rao, J. N. K. (2018). Small area estimation of complex parameters under unit-level models with skew-normal errors. *Scandinavian Journal of Statistics*, *45*, 1092–1116.

Dunnett, C. W., & Sobel, M. (1955). Approximations to the probability integral and certain percentage points of a multivariate analogue of Student's *t*-distribution. *Biometrika*, *42*, 258–260.

Ferraz, V. R. S., & Moura, F. A. S. (2012). Small area estimation using skew normal models. *Computational Statistics and Data Analysis*, *56*, 2864–2874.

Fuller, W. A., & Battese, G. E. (1973). Transformations for estimation of linear models with nested-error structure. *Journal of the American Statistical Association*, *68*, 626–632.

Ghosh, M., & Rao, J. N. K. (1994). Small area estimation: an appraisal. *Statistical Science*, *9*, 55–76.

Henze, N. (1986). A probabilistic representation of the "skew-normal" distribution. *Scandinavian Journal of Statistics*, *13*, 271–275.

Pewsey, A. (2000). Problems of inference for Azzalini's skew-normal distribution. *Journal of Applied Statistics*, *27*, 859–870.

Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science*, *28*, 40–68.

Prasad, N. G. N., & Rao, J. N. K. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, *85*, 163–171.

Rao, J. N. K., & Molina, I. (2015). *Small area estimation* (2nd ed.). Hoboken: Wiley.

Tallis, G. M. (1961). The moment generating function of the truncated multi-normal distribution. *Journal of the Royal Statistical Society: Series B*, *23*, 223–229.