博士論文

Mixed Modeling Approach to Small Area Estimation

（混合効果モデルによる小地域推定へのアプローチ）

Shonosuke SUGASAWA

菅澤翔之助

Contents

Abstract

When area-wise sample sizes are small, the direct estimators of area-specific parameters based only on the samples within each area are unstable. Hence, we need to "borrow strength" across related areas to produce reliable indirect (model-based) estimators of the area-specific parameters, which is known as small area estimation (SAE). Typically, regression models with random effects (mixed models) are used and various mixed methods for SAE have been developed so far. However, most existing models are not flexible enough to capture complex characteristics of data, which might lead to inefficient model-based estimators. This thesis develops several mixed modeling approaches for SAE to overcome the problem.

Chapter 1 briefly explains backgrounds and motivations of the works given in the subsequent chapters. Chapter 2 and 3 propose the use of a parametric family of transformations for response values of observed data. Chapter 4 deals with conditional mean squared errors for risk evaluation of model-based estimators. Chapter 5 develops a small area model with heteroscedastic variances expressed as a function of covariates. Chapter 6 proposes a method for shrinkage estimation of area means as well as sampling variances. Chapter 7 and 8 develop mixed models with uncertain random effects whose distribution is expressed as a mixture of one point distribution and a continuous distribution.

Chapter 1

Introduction

Small area estimation (SAE) deals with the problem of producing reliable estimates of area-specific parameters. Direct estimates based only on the area-specific sample data are not suitable when the sample size is not large. Hence, we need to "borrow strength" across related areas to produce reliable indirect (model-based) estimates for small areas. In SAE, mixed models have been widely used for variety purposes. For comprehensive overviews and appraisals of models and methods for SAE, see Pfeffermann (2013) and Rao and Molina (2015).

The mixed models for SAE can be divided into two major parts, area-level models and unit-level models. In area-level models, the Fay-Heriot (FH) model (Fay and Herriot, 1979) is most famous and extensively used as the standard tool for SAE of continuous valued parameters. When observed values are count or binary, generalized linear mixed models (Jiang, 2006) or models based on natural exponential family with conjugate priors (Ghosh and Maiti, 2004) are useful alternatives. On the other hand, for unit-level data, the nested error regression (NER) model (Battese et al, 1988) is widely used. In these models, random effects play an crucial roles representing the difference between small areas, and the prediction (estimation) of random effects is a key to SAE. Although these mixed models generally perform well and are easy to fit, these models may produce inefficient and biased small area estimates when data does not satisfy assumptions in these models. To overcome this problem, this thesis proposes alternative mixed modeling approaches for SAE.

Chapter 2 addresses the problem of transforming response values in the FH model. In many applications, response values take positive values (e.g. income, consumption) and the distribution is often skewed while the response values in the FH model are assumed to be normal. This inconsistency could cause a considerable bias in the resulting small area estimator. Hence, we propose the use of a parametric family of transformations and generalize the results obtained in the FH model. We derive the empirical best predictor of the small area parameter and a second-order unbiased estimator of the mean squared error of the predictor based on parametric bootstrap. We assess the approach via simulations and an application to survey of family income and expenditure (SFIE) in Japan. This chapter comes from the paper of Sugasawa and Kubokawa (2015) and Sugasawa and Kubokawa (2017b).

Chapter 3 deals with the problem of estimating finite population parameters based on partially observed units. Concerning this problem, Molina and Rao (2010) suggested an empirical best prediction approach based on the NER model. Molina and Rao (2010) applied

their method to estimating area-specific poverty indicators based on unit level income data. Since the income data is skewed, Molina and Rao (2010) used log-transformation before fitting the NER model. However, if the log-transformation is misspecified, the predicted values from the empirical best prediction method are not reliable. Hence, similarly to Chapter 2, we suggest the use of a parametric family of transformations for flexible prediction of a finite population parameter. We sketch a simple estimating method of the model parameters including transformation parameters, and derive transformed empirical best predictors. We compare the proposed method with the method by Molina and Rao (2010) through simulations and an application to synthetic income data in Spanish provinces. This chapter comes from the paper of Sugasawa and Kubokawa (2017d).

Chapter 4 discusses a new risk measure for small area estimators, conditional mean squared errors (CMSE). Traditionally, for measuring the variability of small area estimators, (unconditional) mean squared errors (MSE) have been used. However, as discussed in Booth and Hobert (1998), Datta et al. (2011a), CMSE is more preferable than MSE in the context of small area estimation. Until now, it has been revealed that CMSE and MSE are asymptotically equivalent in small area models based on normal distributions while the difference is not negligible under non-normality. We investigate CMSE in the models based on natural exponential family with quadratic variance function developed by Ghosh and Maiti (2004). We also derive a second-order unbiased estimator of CMSE and show the difference between CMSE and MSE through applications to stomach cancer data and infant mortality data. The result in this chapter was published in Sugasawa and Kubokawa (2016).

Chapter 5 deals with a problem regarding heteroscedastic variances in the NER model. While the NER model assumes that all units are homoscedastic, Jiang and Nguyen (2012) demonstrated that such a structure is restrictive in practice and may produce inefficient estimates. To solve this problem, we propose a heteroscedastic NER model in which the heteroscedastic variances are represented by smooth parametric functions of covariates. We propose a moment method for estimating model parameters and derive an empirical best linear unbiased predictor of the small area parameter. We assess the approach via simulations and an application to posted land price data. This chapter comes from the paper of Sugasawa and Kubokawa (2017a).

Chapter 6 tackles the problem of estimating sampling variances in the FH model. In the conventional FH model, the sampling variance of the direct estimator is assumed to be known while the estimated sampling variances are used in practice. However, it has been recognized that the model-based estimator could produce poor estimates when the estimated sampling variances are unstable. We propose a hierarchical model which produces shrinkage estimators of means as well as variances. We sketch an efficient computational method relying on Markov Chain Monte Carlo (MCMC) and evaluate the proposed model through simulations and empirical studies of SFIE data and famous crop data. This chapter comes from the paper of Sugasawa et al. (2017a).

Chapter 7 proposes the use of the uncertain random effect whose distribution is expressed as a mixture distribution of normal and a point mass on 0, in the NER model. Datta and Mandal (2015) showed that the use of uncertain random effects can substantially improve the estimation accuracy of the model-based estimators. However, their method is restrict to the Fay-Herriot model. We consider using the ideal of uncertain random effects in the NER model. We develop a MCMC method based on Gibbs sampling for computing small area estimators as well as estimates of model parameters. We compare the proposed method with

the traditional NER model via simulations and an application to posted land price data. This chapter comes from the paper of Sugasawa and Kubokawa (2017c).

Chapter 8 deals with the uncertainty of random effects in the context of models based on natural exponential family (Ghosh and Maiti, 2004). We suggest a mixture prior of the conjugate prior and a point mass. Due to the conjugacy of the prior, an Expectation-Maximization (EM) algorithm for estimating model parameters can be easily implemented. Then, we propose an empirical uncertain Bayes estimator and also provide a second order unbiased estimator of CMSE for risk evaluation. The performances of the proposed method are evaluated via simulations and applications to historical mortality data and poverty data in Spain. The content in this chapter comes from Sugasawa et al. (2017b).

Chapter 2

Transforming Responce Values in Fay-Herriot Model

2.1 Introduction

The basic random effect model for area-level data is the Fay-Herriot (FH) model (Fay and Herriot, 1979). Let θ_i be the small area mean (or total) in the ith area and let y_i denotes the direct estimator of θ_i. The FH model is defined as

$$y_i = \theta_i + \varepsilon_i, \quad \theta_i = \boldsymbol{x}_i^t \boldsymbol{\beta} + v_i, \quad i = 1, \ldots, m, \tag{2.1}$$

where $\boldsymbol{x}_i$ and $\boldsymbol{\beta}$ are p-dimensional vectors of covariates and regression coefficients, respectively, and v_i and ε_i are mutually independent and distributed as $v_i \sim N(0, A)$ for unknown variance parameter A and $\varepsilon_i \sim N(0, D_i)$ for known sampling variance D_i. The known variance D_i is typically obtained by smoothing the sampling variance and then treating the smoothed estimates as the true D_i (Rao and Molina 2015). Under squared error loss, the Bayes estimator of θ_i is obtained as

$$\widetilde{\theta}_i = \gamma_i y_i + (1 - \gamma_i) \boldsymbol{x}_i^t \boldsymbol{\beta},$$

where $\gamma_i = A/(A + D_i)$. It is observed that the Bayes estimator $\widetilde{\theta}_i$ is the weighted linear combination of the direct estimator y_i and the synthetic estimator $\boldsymbol{x}_i^t \boldsymbol{\beta}$. Since the model parameters $\boldsymbol{\beta}$ and A are unknown, we estimate them from the data $\{(y_i, \boldsymbol{x}_i),\ i = 1, \ldots, m\}$. The generalized least squares estimator is typically used for $\boldsymbol{\beta}$ while several methods, including the (restricted) maximum likelihood estimator, are used for estimating A.

In the Fay-Herriot (FH) model (2.1), it is assumed that $y_i \sim N(\boldsymbol{x}_i^t \boldsymbol{\beta}, A + D_i)$, namely the response variable (direct estimator) y_i is normally distributed. However, we often encounter positive-valued data (e.g. income, expense), which have skewed distributions and non-linear relationships with covariates. For such a data set, the traditional FH model with a linear structure between direct estimates and covariates and normally distributed error terms is not clearly appropriate. A common approach is using the log-transformed direct estimators and apply the FH model (e.g. Slud and Maiti, 2006). However, the log-transformation is not always appropriate and it may produce inefficient and biased prediction when the log-transformation is misspecified. Thus, a more natural approach to tackle this issue is using a parametric family of transformations which enables us to flexibly select a reasonable

transformation based on data. A famous family is the Box-Cox transformation (Box and Cox, 1964), but it is well-known that the Box-Cox transformation suffers from the truncation problem, which leads to inconsistency of the maximum likelihood estimator of λ, and the inverse transformation can not be defined on whole real line, so that we can not drive a back-transformed predictors in the original scale. Thus the use of the Box-Cox transformation in the context of small area estimation is not desirable. Instead of the Box-Cox transformation, Yang (2006) suggested a novel family of transformations called the dual power (DP) transformation

$$h_\lambda(x) = \begin{cases} (2\lambda)^{-1}(x^\lambda - x^{-\lambda}) & \lambda > 0 \\ \log x & \lambda = 0, \end{cases}$$

which can be seen as the average of two Box-Cox transformations. The main advantage of the DPT is that its range is the whole real line for all $\lambda \geq 0$, and it does not suffer from the truncation problem. Hence, the use of the DP transformation in the FH model seems an attractive approach for estimating positive valued small area parameters. This chapter introduces a new transformation approach to the FH model with the DP transformation. In Section 2.2, we describe the proposed model and provide methods for parameter estimation and computing small are estimators. A second-order unbiased estimator of mean square errors (MSE) of small area estimators are derived for measuring the variability of small area estimators. In Sections 2.3 and 2.4, we present some simulation studies and empirical applications, respectively. The technical details are given in Section 2.5.

2.2 Transformed Fay-Herriot Model

2.2.1 Model setup and best predictor

We consider the following parametric transformed Fay-Herriot (PTFH) model for area-level data:

$$h_\lambda(y_i) = \boldsymbol{x}_i^t\boldsymbol{\beta} + v_i + \varepsilon_i, \quad i = 1, \ldots, m \tag{2.2}$$

where $v_i \sim N(0, A)$, $\varepsilon_i \sim N(0, D_i)$ for known D_i's, $\boldsymbol{\beta}$ and $\boldsymbol{x}_i$ are p-dimensional vectors of regression coefficients and covariates, respectively. The unknown models parameters are denoted by $\boldsymbol{\phi} = (\boldsymbol{\beta}^t, A, \lambda)^t$ and we aim to estimate (predict) $\mu_i = h_\lambda^{-1}(\theta_i)$ with $\theta_i = \boldsymbol{x}_i^t\boldsymbol{\beta} + v_i$. Note that when $\lambda = 0$, the model (2.2) reduces to the log-transformed Fay-Herriot model studied by Slud and Maiti (2006).

It is well known that the best predictor of θ_i under the squared error loss is given by

$$\widetilde{\theta}_i = \gamma_i h_\lambda(y_i) + (1 - \gamma_i)\boldsymbol{x}_i^t\boldsymbol{\beta}, \tag{2.3}$$

where $\gamma_i = A/(A + D_i)$. Hence, one possible way to predict μ_i is using the simple back-transformed predictor $\widetilde{\mu}_i^{(S)} = h_\lambda^{-1}(\widetilde{\theta}_i)$. However, $\widetilde{\mu}_i^{(S)}$ is not suitable for predicting μ_i, because $\widetilde{\mu}_i^{(S)}$ has a non-ignorable bias for predicting μ_i, namely $\mathrm{E}[\widetilde{\mu}_i^{(S)} - \mu_i] \neq 0$ even when m is large. On the other hand, Slud and Maiti (2006) considered the bias corrected predictor of $F(\theta_i)$ for a general function $F(\cdot)$, which leads to the following form:

$$\widetilde{\mu}_i^{(SM)} = \frac{\mathrm{E}[h_\lambda^{-1}(\theta_i)]}{\mathrm{E}[h_\lambda^{-1}(\widetilde{\theta}_i)]} h_\lambda(\widetilde{\theta}_i) = \frac{\mathrm{E}[\mu_i]}{\mathrm{E}[\widetilde{\mu}_i^{(S)}]} \widetilde{\mu}_i^{(S)}.$$

It clearly holds that $\mathrm{E}[\widetilde{\mu}_i^{(SM)} - \mu_i] = 0$, that is, $\widetilde{\mu}_i^{(SM)}$ is an unbiased predictor of μ_i while it does not necessarily minimize the squared error loss. We here use the conditional expectation $\widetilde{\mu}_i = \mathrm{E}[h_\lambda^{-1}(\theta_i)|y_i]$ with known $\boldsymbol{\phi}$ as a predictor of μ_i, which minimizes the squared error loss. Since $\theta_i|y_i \sim N(\widetilde{\theta}_i, \sigma_i^2)$ with $\widetilde{\theta}_i$ given in (3.5) and $\sigma_i^2 = AD_i/(A+D_i)$ under the model (2.2), the conditional expectation $\widetilde{\mu}_i$ can be expressed as

$$\widetilde{\mu}_i \equiv \widetilde{\mu}_i(y_i; \boldsymbol{\phi}) = \int_{-\infty}^{\infty} h_\lambda^{-1}(t)\phi(t; \widetilde{\theta}_i, \sigma_i^2)dt, \tag{2.4}$$

where $\phi(\cdot; a, b)$ denotes the density function of $N(a, b)$. It should be noted that $\widetilde{\mu}_i = \widetilde{\mu}_i^{(SM)}$ when $\lambda = 0$, namely $h_\lambda^{-1}(x) = \exp(x)$. However, $\widetilde{\mu}_i$ and $\widetilde{\mu}_i^{(SM)}$ are not necessarily identical when $\lambda > 0$.

Since the model parameters $\boldsymbol{\phi}$ is unknown in practice, we estimate them by maximizing the marginal likelihood function, and the details are given in the next section. Let $\widehat{\boldsymbol{\phi}}$ be the corresponding estimator of $\boldsymbol{\phi}$. Then, replacing $\boldsymbol{\phi}$ with $\widehat{\boldsymbol{\phi}}$ in (2.4) leads to the empirical form of $\widetilde{\mu}_i$:

$$\widehat{\mu}_i \equiv \widetilde{\mu}(y_i; \widehat{\boldsymbol{\phi}}) = \int_{-\infty}^{\infty} h_{\widehat{\lambda}}^{-1}(t)\phi(t; \widehat{\theta}_i, \widehat{\sigma}_i^2)dt,$$

which is known as empirical best predictor (EBP). Note that $\widehat{\mu}_i$ is no longer the conditional expectation but $\widehat{\mu}_i$ converges to $\widetilde{\mu}_i$ as $m \to \infty$ under some regularity conditions. Since $\widehat{\mu}_i$ cannot be obtained in an analytical form, we rely on numerical techniques for computing $\widehat{\mu}_i$. A typical method is the Monte Carlo integration by generating a large numbers of random samples from $(\widehat{\theta}_i, \widehat{\sigma}_i^2)$. However, we here use Gaussian-Hermite quadrature which is known to be more accurate than the Monte Carlo integration.

2.2.2 Estimation of model parameters

Under normality assumption of v_i and ε_i, it follows that $h_\lambda(y_i) \sim N(\boldsymbol{x}_i^t\boldsymbol{\beta}, A + D_i)$ and $h_\lambda(y_i),\ i = 1, \ldots, m$ are mutually independent. Then, the maximum likelihood estimator $\widehat{\boldsymbol{\phi}}$ of $\boldsymbol{\phi}$ is defined as the maximizer of $L(\boldsymbol{\phi})$, where

$$L(\boldsymbol{\phi}) = -\sum_{i=1}^{m} \log(A + D_i) - \sum_{i=1}^{m} \frac{\left\{h_\lambda(y_i) - \boldsymbol{x}_i^t\boldsymbol{\beta}\right\}^2}{A + D_i} + 2\sum_{i=1}^{m} \log\left(y_i^{\lambda-1} + y_i^{-\lambda-1}\right). \tag{2.5}$$

Note that the third term in (2.5) comes from the Jacobian of the transformation. When λ is given, maximizing (2.5) with respect to $\boldsymbol{\beta}$ and A coincides to maximizing the log-likelihood function of the classical Fay-Herriot model. Hence, the value of profile likelihood function of λ is easily computed, so that we may estimate λ by grid search over a specified region or golden section method (Brent, 1973). Though the parameter space of λ is $[0, \infty)$, it would be sufficient to consider the space $[0, \lambda_m]$ for moderately large λ_m.

For asymptotic properties of the estimator $\widehat{\boldsymbol{\phi}}$, we assume the following conditions.

Assumption 2.1.

1. *There exist* $\underline{D}$ *and* $\overline{D}$ *independent to* m *such that* $\underline{D} \le D_i \le \overline{D}$ *for* $i = 1, \ldots, m$.
2. $\max_{i=1,\ldots,m} \boldsymbol{x}_i^t(\sum_{j=1}^m \boldsymbol{x}_j\boldsymbol{x}_j^t)^{-1}\boldsymbol{x}_i = O(m^{-1})$.

These conditions are usually assumed in the context of small area estimation, see Datta and Lahiri (2000) and Butar and Lahiri (2003). Under these conditions, we have the following lemma.

Lemma 2.1. *Under Assumption 2.1, as* $m \to \infty$, $\sqrt{m}(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi})$ *asymptotically follows the multivariate normal distribution* $N(\mathbf{0}, \boldsymbol{V}(\boldsymbol{\phi}))$ *with a covariance matrix* $\boldsymbol{V}(\boldsymbol{\phi})$*, and it holds* $\mathrm{E}[\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}] = m^{-1}\boldsymbol{b}(\boldsymbol{\phi}) + o(m^{-1})$ *with a smooth function* $\boldsymbol{b}(\boldsymbol{\phi})$*.*

The asymptotic normality of $\widehat{\boldsymbol{\phi}}$ immediately follows from Sugasawa and Kubokawa (2015). Moreover, from the proof of Theorem 1 in Lohr and Rao (2009), the bias $\boldsymbol{b}(\boldsymbol{\phi})$ can be expressed by partial derivatives of $L(\boldsymbol{\phi})$ given in (2.5), so that the latter statement in Lemma 8.1 follows.

Other estimators of A are the restricted maximum likelihood estimator (Jiang, 1996), the Prasad-Rao estimator (Prasad and Rao, 1990), the Fay-Herriot estimator (Fay and Herriot, 1979) and the adjusted maximum likelihood estimator (Li and Lahiri, 2010). These methods can be easily implemented and their asymptotic properties are discussed in Sugasawa and Kubokawa (2015). However, for simplicity, we do not treat these estimators in this paper.

2.2.3 Mean squared error of the empirical best predictor

In small area estimation, mean squared errors (MSEs) of small area estimators are used for risk evaluation, and their importance has been addressed in many papers including Lahiri and Rao (1995) and Datta et al. (2005). Following this convention, we evaluate the MSE of the empirical best predictor $\widehat{\mu}_i$. To begin with, we note that the MSE can be decomposed as

$$\begin{aligned}\mathrm{MSE}_i \equiv \mathrm{E}[(\widehat{\mu}_i - \mu_i)^2] &= \mathrm{E}[(\widetilde{\mu}_i - \mu_i)^2] + \mathrm{E}[(\widehat{\mu}_i - \widetilde{\mu}_i)^2] \\ &\equiv g_{1i}(\boldsymbol{\phi}) + g_{2i}(\boldsymbol{\phi}),\end{aligned}$$

because $\widetilde{\mu}_i = E[\mu_i|y_i]$ is the conditional expectation. In what follows, we use the explicit notation $\widetilde{\mu}_i(y_i, \boldsymbol{\phi})$ instead of $\widetilde{\mu}_i$ if necessary. The first term $g_{1i}(\boldsymbol{\phi})$ is expressed as

$$g_{1i}(\boldsymbol{\phi}) = \mathrm{E}\left[\left\{\widetilde{\mu}_i(\boldsymbol{x}_i^t\boldsymbol{\beta} + v_i + \varepsilon_i, \boldsymbol{\phi}) - h_\lambda^{-1}(\boldsymbol{x}_i^t\boldsymbol{\beta} + v_i)\right\}^2\right],$$

which has no analytical expression. The direct Monte Carlo integration by generating random samples of v_i and ε_i requires a large computational burden because we need another Monte Carlo integration for computing $\widetilde{\mu}_i$ for each sample (v_i, ε_i). However, as shown in the Appendix, it turns out to have the following more simple expression of $g_{1i}(\boldsymbol{\phi})$:

$$g_{1i}(\boldsymbol{\phi}) = \mathrm{E}\left[\{h_\lambda^{-1}(\boldsymbol{x}_i^t\boldsymbol{\beta} + z_1)\}^2 - h_\lambda^{-1}(\boldsymbol{x}_i^t\boldsymbol{\beta} + c_{1i}z_1 + c_{2i}z_2)h_\lambda^{-1}(\boldsymbol{x}_i^t\boldsymbol{\beta} + c_{1i}z_1 - c_{2i}z_2)\right], \quad (2.6)$$

where $z_1, z_2 \sim N(0, A)$, $c_{1i} = \sqrt{(1+a_i)/2}$, and $c_{2i} = \sqrt{(1-a_i)/2}$ for $a_i = A/(A+D_i)$. Hence, $g_{1i}(\boldsymbol{\phi})$ can be easily calculated by generating a large number of random samples of z_1 and z_2. On the other hand, the second term $g_{2i}(\boldsymbol{\phi})$ can be evaluated as the following lemma, where the proof is given in the Appendix.

Lemma 2.2. *Under Assumption 2.1, it holds*

$$g_{2i}(\boldsymbol{\phi}) = \frac{1}{m}\mathrm{tr}\left\{\boldsymbol{V}(\boldsymbol{\phi})\mathrm{E}\left[\frac{\partial\widetilde{\mu}_i}{\partial\boldsymbol{\phi}}\frac{\partial\widetilde{\mu}_i}{\partial\boldsymbol{\phi}^t}\right]\right\} + o(m^{-1}).$$

Since the MSE depends on unknown parameter $\boldsymbol{\phi}$, we need to estimate it for practical use. To this end, we obtain a second-order unbiased estimator of the MSE. Here, an estimator $\widehat{B}$ is called second order unbiased if $\mathrm{E}[\widehat{B}] = B + o(m^{-1})$. From lemma 2.2, it follows that $g_{2i}(\boldsymbol{\phi}) = m^{-1}c_1(\boldsymbol{\phi}) + o(m^{-1})$ with the smooth function $c_1(\boldsymbol{\phi})$, thereby the plug-in estimator $g_{2i}(\widehat{\boldsymbol{\phi}})$ is second-order unbiased. However, the plug-in estimator $g_{1i}(\widehat{\boldsymbol{\phi}})$ has a second-order bias since $g_{1i}(\boldsymbol{\phi}) = O(1)$, so that we need to correct the bias. Hence, we propose the parametric bootstrap method to correct the bias of $g_{1i}(\widehat{\boldsymbol{\phi}})$ and computing $g_{2i}(\widehat{\boldsymbol{\phi}})$. The procedure is given in the following.

Parametric bootstrap method for the MSE estimation

1. Generate bootstrap samples y_i^* from the estimated model;

$$h_{\widehat{\lambda}}(y_i^*) = \boldsymbol{x}_i^t\widehat{\boldsymbol{\beta}} + v_i^* + \varepsilon_i^*, \quad i = 1, \ldots, m, \tag{2.7}$$

   where ε_i^* and v_i^* are generated from $N(0, D_i)$ and $N(0, \widehat{A})$, respectively

2. Based on $(y_i^*, \boldsymbol{x}_i),\ i = 1, \ldots, m$, compute the maximum likelihood estimate $\widehat{\boldsymbol{\phi}}^*$ and the predicted values of $\widehat{\mu}_i = \widetilde{\mu}_i(y_i, \widehat{\boldsymbol{\phi}})$ and $\widehat{\mu}_i^* = \widetilde{\mu}_i(y_i, \widehat{\boldsymbol{\phi}}^*)$.

3. Derive the bootstrap estimates of g_{1i} and g_{2i} via

$$g_{1i}^{\mathrm{bc}}(\widehat{\boldsymbol{\phi}}) = 2g_{1i}(\widehat{\boldsymbol{\phi}}) - E^*\left[g_{1i}(\widehat{\boldsymbol{\phi}}^*)\right], \quad g_{2i}^*(\widehat{\boldsymbol{\phi}}) = E^*\left[(\widehat{\mu}_i^* - \widehat{\mu}_i)^2\right]$$

   where $\widehat{\mu}_i^* = h_{\widehat{\lambda}}^{-1}(\boldsymbol{x}_i^t\widehat{\boldsymbol{\beta}} + v_i^*)$ and $E^*[\cdot]$ denotes the expectation with respect to the bootstrap samples generated from (2.7). The second-order unbiased estimator of the MSE based on the parametric bootstrap is given by

$$\widehat{\mathrm{MSE}_i} = g_{1i}^{\mathrm{bc}}(\widehat{\boldsymbol{\phi}}) + g_{2i}^*(\widehat{\boldsymbol{\phi}}). \tag{2.8}$$

The resulting MSE estimator (2.8) is second-order unbiased as shown in the following theorem, which is proved in the Appendix.

Theorem 2.1. *Let* $\widehat{\mathrm{MSE}_i}$ *be the parametric bootstrap MSE estimator given in* (2.8)*. Then, under Assumption 2.1, we have*

$$\mathrm{E}\left[\widehat{\mathrm{MSE}_i}\right] = \mathrm{MSE}_i + o(m^{-1}),$$

where the expectation is taken with respect to y_i*'s following the model (2.2).*

In (2.8), the bias correction of $g_{1i}(\widehat{\boldsymbol{\phi}})$ is carried out via using the additive form $g_{1i}^{\mathrm{bc}}(\widehat{\boldsymbol{\phi}}) = 2g_{1i}(\widehat{\boldsymbol{\phi}}) - E^*[g_{1i}(\widehat{\boldsymbol{\phi}}^*)]$, where E^* denotes the expectation with respect to bootstrap samples. Hall and Maiti (2006a) suggested other bias-correcting methods including a multiplicative bias correcting method of the form $g_{1i}(\widehat{\boldsymbol{\phi}})^2/E^*[g_{1i}(\widehat{\boldsymbol{\phi}}^*)]$. The multiplicative form for bias correction can avoid negative estimates of the MSE while the additive form for bias correction gives negative estimates of the MSE with a positive probability. Although those bias corrections give second-order unbiased estimates of $g_{1i}(\boldsymbol{\phi})$, in this paper, we use the additive-type bias correction, because it has been frequently used in the literatures (e.g. Butar and Lahiri, 2003).

2.3 Simulation Studies

2.3.1 Evaluation of prediction errors

We evaluated prediction errors of the proposed PTFH model and some existing models. As a data generating process, we considered the following PTFH model:

$$h_\lambda(y_i) = \beta_0 + \beta_1 x_i + v_i + \varepsilon_i, \quad i = 1, \ldots, 30, \tag{2.9}$$

where $v_i \sim N(0, A)$, $\varepsilon_i \sim N(0, D_i)$ with $\beta_0 = 1, \beta_1 = 1$ and $A = 1.5$. For λ, we treated the four cases $\lambda = 0.1, 0.4, 0.7$ and 1.0. The covariates x_i were initially generated from the uniform distribution on $(0, 4)$ and fixed in simulation runs. Concerning sampling variance D_i, we divided 30 areas into 5 groups (from G_1 to G_5), and areas within the same group have the same D_i value. The D_i-pattern we considered was $(0.2, 0.4, 0.6, 0.8, 1.0)$. The true small area parameters are $\mu_i = h_\lambda^{-1}(\beta_0 + \beta_1 x_i + v_i)$.

For comparison, we considered the log-transformed FH (log-FH) model and the traditional Fay-Herriot (FH) model, which are described as

$$\begin{aligned} \text{log-FH:} \quad & \log y_i = \beta_0 + \beta_1 x_i + v_i + \varepsilon_i \\ \text{FH:} \quad & y_i = \beta_0 + \beta_1 x_i + v_i + \varepsilon_i. \end{aligned}$$

It is noted that the data generating process (2.9) get close to log-FH as λ gets smaller.

Since we do not known the true D_i in practice, we computed the estimates of μ_i with estimated D_i as investigated in Bell (2008). To this end, we generated the auxiliary observation z_{ik} from the model:

$$h_\lambda(z_{ik}) = \varepsilon_{ik}, \quad i = 1, \ldots, 30, \quad k = 1, \ldots, 10, \tag{2.10}$$

where $\varepsilon_{ij} \sim N(0, D_i)$. In applying log-FH and FH, we computed the estimates of D_i as the sampling variances of $\{\log z_{i1}, \ldots, \log z_{i10}\}$ and $\{z_{i1}, \ldots, z_{i10}\}$, respectively. Then we computed the estimates of μ_i using EBLUP in FH and the bias-corrected estimator used in Slud and Maiti (2006) in log-FH, where the model parameters β_0, β_1 and A are estimated via the maximum likelihood method. For fitting PTFH, we first define

$$D_i \equiv D_i(\lambda) = \frac{1}{9}\sum_{k=1}^{10}\left\{h_\lambda(z_{ik}) - \overline{h_\lambda(z)}_i\right\}^2, \quad \overline{h_\lambda(z)}_i = \frac{1}{10}\sum_{k=1}^{10} h_\lambda(z_{ik}),$$

so that we regard D_i as a function of λ and replace D_i with $D_i(\lambda)$ in (2.5). Since $D_i(\lambda)$ can be immediately computed under the given λ, we can maximize the profile likelihood function of λ in the similar manner to that presented in Section 2.2.2. Once the estimate $\widehat{\lambda}$ is computed, D_i can be calculated as $D_i(\widehat{\lambda})$.

Based on $R = 10000$ simulation runs, we computed the coefficient of variation (CV) and the absolute relative bias (ARB), defined as

$$\mathrm{CV}_i = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\frac{\left(\widehat{\mu}_i^{(r)} - \mu_i^{(r)}\right)^2}{\mu_i^{(r)2}}} \quad \text{and} \quad \mathrm{ARB}_i = \left|\frac{1}{R}\sum_{r=1}^{R}\frac{\widehat{\mu}_i^{(r)} - \mu_i^{(r)}}{\mu_i^{(r)}}\right|,$$

where $\mu_i^{(r)}$ is the true value and $\widehat{\mu}^{(r)}$ is the estimated value from PTFH, log-FH or FH, in the rth iteration. Table 8.1 shows the percent CV and ARB averaged within the same groups

for each case of λ. For comparison, we also show the results for PTFH with true D_i values, denoted by PTFH-t in Table 8.1.

From Table 8.1, we can observe that difference CV or ARB between PTFH-t and PTFH tends to be large when λ is small and D_i is large while tow methods perform similarly when λ is large or D_i is small. Moreover, it is revealed that the larger D_i would lead to larger CV and ARB values in all the methods. Concerning comparison among PTFH, log-FH and FH, it can be seen that PTFH performs better than FH except for $\lambda = 0.9$, and PTFH performs better than log-FH except for $\lambda = 0.1$. Moreover, the differences between PTFH and log-FH in $\lambda = 0.1$ get larger as D_i gets larger. Regarding PTFH-t, it performs best in most cases. However, it is observed that log-FH and FH produce more accurate estimates than PTFH-t in some cases.

We next investigated the prediction errors when the true distribution of v_i is not normal. Here, we considered a t-distribution with 5 degrees of freedom for v_i, where the variance is scaled to A, and the other settings for the data generation are the same as (2.9). Under the scenario, we again computed the values of CV and ARB of the four methods based on 10000 simulation runs, and the results are reported in Table 2.2. It is observed that the simulated RMSE values in Table 2.2 are larger than those in Table 8.1 due to misspecification of the distribution of v_i. However, relationships of CV and ARB among three methods are similar to Table 8.1.

2.3.2 Finite sample performance of the MSE estimator

We next investigated a finite sample performance of the MSE estimator (2.8). Following Datta et al. (2005), we considered the following data generating process without covariates:

$$h_\lambda(y_i) = \mu + v_i + \varepsilon_i, \qquad i = 1, \ldots, 30,$$

with $\mu = 0$, $v_i \sim N(0, A)$ with $A = 1$ and $\varepsilon_i \sim N(0, D_i)$. As a value of λ, we considered the three cases $\lambda = 0.2, 0.6, 1.0$. For setting of D_i, we divided D_i's into five groups $G_1, \ldots, G_5$, where D_i's were the same values over the same group, and the following three patterns of D_i's were considered:

$$(a)\ 0.3, 0.4, 0.5, 0.6, 0.7, \quad (b)\ 0.2, 0.4, 0.5, 0.6, 2.0, \quad (c)\ 0.1, 0.4, 0.5, 0.6, 4.0.$$

Based on $R_1 = 5000$ simulation runs, we calculated the simulated values of the MSE as

$$\mathrm{MSE}_i = \frac{1}{R_1}\sum_{r=1}^{R_1}(\widehat{\mu}_i^{(r)} - \mu_i^{(r)})^2, \qquad \mu_i^{(r)} = h_\lambda^{-1}(\beta_0 + v_i^{(r)})$$

where $\widehat{\mu}_i^{(r)}$ and $v_i^{(r)}$ are the predicted value and the realized value of v_i in the r-th iteration. Then based on $R_2 = 2000$ simulation runs, we calculated the relative bias (RB) and the coefficient of variation (CV) defined as

$$\mathrm{RB}_i = \frac{1}{R_2}\sum_{r=1}^{R_2}\left(\widehat{\mathrm{MSE}_i}^{(r)} - \mathrm{MSE}_i\right)/\mathrm{MSE}_i,$$

$$\mathrm{CV}_i^2 = \frac{1}{R_2}\sum_{r=1}^{R_2}\left(\widehat{\mathrm{MSE}_i}^{(r)} - \mathrm{MSE}_i\right)^2/\mathrm{MSE}_i^2.$$

Table 2.1: Simulated percent coefficient of variation (CV) and absolute relative biases (ARB) of the parametric transformed Fay-Herriot with use of true D_i (PTFH-t) and estimated D_i (PTFH), the log-transformed Fay-Herriot (log-FH) model, and the Fay-Herriot (FH) model under $\lambda = 0.1, 0.4, 0.7$ and 1.0.

		CV				ARB			
	method	0.1	0.4	0.7	1.0	0.1	0.4	0.7	1.0
G_1	PTFH-t	46.49	33.28	23.49	17.95	13.67	6.89	3.69	2.43
	PTFH	47.46	32.90	23.12	17.53	13.97	6.89	3.51	2.21
	log-FH	47.11	35.34	28.63	23.57	13.16	4.18	3.85	2.39
	FH	50.24	32.94	22.84	16.87	9.74	3.82	1.78	1.13
G_2	PTFH-t	63.92	47.64	37.09	31.61	20.20	11.85	8.13	6.69
	PTFH	66.32	47.84	36.92	31.38	21.28	12.24	8.31	6.74
	log-FH	66.79	53.25	46.70	41.76	22.25	15.14	13.96	13.86
	FH	82.25	52.67	38.61	31.11	19.95	9.16	6.27	5.98
G_3	PTFH-t	75.92	59.42	48.51	40.60	25.67	16.79	12.85	9.59
	PTFH	79.86	60.35	48.22	40.35	27.67	17.74	13.15	9.67
	log-FH	77.70	60.99	53.25	47.15	26.56	15.97	13.04	11.30
	FH	116.09	74.04	53.70	40.88	30.55	17.10	11.77	9.25
G_4	PTFH-t	86.92	67.82	53.66	44.73	32.29	20.84	14.11	10.82
	PTFH	92.97	69.13	53.38	43.83	35.12	22.13	14.43	10.63
	log-FH	86.81	65.36	53.46	46.36	30.88	14.42	8.29	8.94
	FH	156.12	91.68	61.04	45.61	45.29	25.10	15.15	11.25
G_5	PTFH-t	92.90	72.74	59.86	49.86	33.87	22.99	17.30	13.04
	PTFH	101.81	75.26	60.39	49.54	37.62	24.73	18.07	13.24
	log-FH	95.91	71.73	61.69	53.58	34.91	20.57	15.30	12.87
	FH	198.30	112.23	73.57	53.02	57.58	31.04	19.05	13.93

For calculation of the MSE estimates in each iteration, we used 100 bootstrap replication for the MSE estimator and 10000 Monte Carlo samples for computing g_{1i}. We also investigated the performance of the MSE estimator when we used the estimated sampling variances instead of known D_i. To this end, similarly to the previous section, we generated the auxiliary observation from (2.10), and calculate D_i's using these data in each simulation run. Based on the same number of simulation runs, we computed the values of RB and CV. Table 2.3 and Table 2.4 show the maximum, mean and minimum values of RB and CV within the same group. In both tables, the simulated values of RB and CV of the MSE estimator with estimated D_i are given in the parenthesis. It is seen that the proposed MSE estimator with known D_i provides reasonable estimated values in almost all cases in terms of both RB and CV. On the other hand, the MSE estimator with estimated D_i performs worse than the MSE estimator with known D_i since the former estimator is affected by the variability of estimating D_i. Moreover, it is observed that performances of both MSE estimators get better in the order of Pattern (a), (b) and (c).

Table 2.2: Simulated percentage coefficient of variation (CV) and percentage absolute relative biases (ARB) of the parametric transformed Fay-Herriot with use of true D_i (PTFH-t) and estimated D_i (PTFH), the log-transformed Fay-Herriot (log-FH) model, and the Fay-Herriot (FH) model under $\lambda = 0.1, 0.4, 0.7$ and 1.0, when the distribution of v_i is a t-distribution with 5 degrees of freedom.

		CV				ARB			
	method	0.1	0.4	0.7	1.0	0.1	0.4	0.7	1.0
G_1	PTFH-t	47.19	39.58	34.16	28.94	14.10	9.77	7.31	5.43
	PTFH	48.42	39.84	34.40	28.34	14.70	10.07	7.38	5.18
	log-FH	48.09	41.08	39.98	35.31	14.00	8.14	4.95	5.65
	FH	51.54	41.09	33.00	27.84	10.26	6.92	5.18	4.28
G_2	PTFH-t	66.06	53.12	41.12	35.83	21.41	14.08	8.99	6.92
	PTFH	68.38	53.52	40.86	35.18	22.59	14.78	9.10	6.80
	log-FH	67.79	56.17	46.04	43.20	21.84	13.32	7.32	6.73
	FH	83.45	58.51	41.73	34.32	21.03	12.63	7.67	6.18
G_3	PTFH-t	79.92	62.21	49.91	42.61	25.50	16.81	10.91	8.69
	PTFH	83.75	62.26	47.41	41.62	27.39	17.60	11.05	8.66
	log-FH	83.53	65.45	55.47	52.08	27.40	18.20	12.93	13.22
	FH	121.91	73.92	50.11	41.05	31.49	17.11	10.35	8.25
G_4	PTFH-t	89.82	78.86	61.81	53.21	30.06	21.02	15.63	12.03
	PTFH	98.67	75.36	59.27	51.18	32.81	22.26	15.83	12.01
	log-FH	99.13	81.96	62.46	57.54	31.16	19.98	13.48	11.01
	FH	155.72	108.51	65.22	52.10	44.14	25.07	15.96	12.80
G_5	PTFH-t	104.87	86.53	77.05	71.23	34.72	26.73	22.00	17.04
	PTFH	123.97	86.69	74.22	64.63	38.87	28.84	22.83	17.06
	log-FH	120.17	87.54	75.41	67.96	34.96	22.70	15.61	11.36
	FH	224.09	128.60	87.85	69.25	61.85	40.85	29.11	21.47

2.4 Application to Survey Data in Japan

We consider an application of the proposed method together with some existing methods to the data from the Survey of Family Income and Expenditure (SFIE) in Japan. Especially, we used the data on the spending item 'Health' and 'Education' in the survey in 2014. For the spending item 'Health' and 'Education', the annual average spending data at each capital city of 47 prefectures are available. The estimates are both unreliable since the sample sizes are around 50 for most prefectures. As a covariate, we used data from the National Survey of Family Income and Expenditure (NSFIE) for 47 prefectures. Since NSFIE is based on much larger sample than SFIE, the reported values are more reliable, but this survey has been implemented every five years. Although the joint bivariate modeling of the two items 'Health' and 'Education' would be preferable as proposed in Benavent and Morales (2016), we here consider applying univariate models separately to each item for simplicity. In what follows, y_i and x_i denote the direct estimate (scaled by 1000) from SFIE and the covariate (reliable estimate) from NSFIE, respectively, on the item 'Health' or 'Education'. For each

Table 2.3: The percentage RB values of the MSE estimator with known D_i and unknown D_i (Parenthesis) in each group.

		Pattern (a)			Pattern (b)			Pattern (c)		
	λ	0.2	0.6	1.0	0.2	0.6	1.0	0.2	0.6	1.0
max	G_1	6.3	17.9	19.5	3.9	7.3	9.0	7.2	1.3	2.7
		(28.1)	(44.3)	(51.4)	(98.4)	(80.8)	(82.6)	(113.5)	(47.1)	(44.1)
	G_2	4.9	15.1	17.2	4.4	4.4	5.3	4.1	−3.6	−4.9
		(44.6)	(69.2)	(78.8)	(46.4)	(107.2)	(99.8)	(37.3)	(29.8)	(20.8)
	G_3	12.5	10.4	14.6	−2.8	3.3	4.6	−3.4	−4.1	−5.0
		(32.9)	(40.8)	(47.0)	(24.2)	(12.5)	(10.6)	(11.3)	(37.5)	(29.7)
	G_4	6.4	12.6	17.6	−2.6	6.7	8.9	−5.0	−4.5	−4.8
		(33.8)	(28.3)	(40.2)	(12.1)	(8.8)	(8.0)	(16.1)	(84.4)	(75.7)
	G_5	8.7	12.4	16.3	−1.5	2.8	6.1	−5.3	−4.2	−1.6
		(15.9)	(39.1)	(52.1)	(6.2)	(43.9)	(43.6)	(4.6)	(26.0)	(19.5)
mean	G_1	2.5	7.4	8.6	−1.9	1.4	2.6	−2.7	−3.3	−3.1
		(22.4)	(31.4)	(36.9)	(33.5)	(32.3)	(33.9)	(47.6)	(27.7)	(27.0)
	G_2	−1.0	7.8	9.8	−3.4	−2.0	−2.0	−6.0	−5.9	−7.3
		(21.9)	(33.4)	(39.9)	(15.5)	(21.5)	(18.8)	(20.2)	(4.8)	(−0.6)
	G_3	5.2	1.9	5.2	−6.3	−1.7	−0.4	−6.1	−8.4	−9.1
		(17.7)	(21.1)	(28.7)	(11.2)	(3.6)	(2.4)	(−0.1)	(10.4)	(4.6)
	G_4	3.3	6.4	9.9	−8.0	−0.4	1.5	−7.4	−7.5	−7.9
		(20.4)	(17.7)	(26.9)	(2.2)	(−2.8)	(−2.2)	(−5.3)	(15.6)	(10.9)
	G_5	1.4	5.7	9.9	−6.0	−0.2	2.4	−7.0	−6.8	−6.4
		(11.1)	(22.1)	(32.8)	(−3.0)	(6.4)	(7.2)	(−6.8)	(−1.2)	(−4.4)
min	G_1	−4.9	−3.2	−2.4	−6.9	−5.2	−4.5	−8.5	−6.9	−6.6
		(17.2)	(15.0)	(20.2)	(13.7)	(16.1)	(17.3)	(21.5)	(17.1)	(17.0)
	G_2	−5.8	−0.9	2.8	−7.5	−7.0	−7.5	−9.6	−7.8	−9.9
		(11.5)	(14.2)	(18.6)	(2.7)	(−2.7)	(−4.7)	(8.1)	(−5.2)	(−8.8)
	G_3	0.9	−8.2	−6.7	−8.7	−6.6	−6.5	−10.3	−10.4	−11.3
		(7.0)	(13.2)	(19.4)	(−5.0)	(−4.4)	(−4.7)	(−7.3)	(−9.4)	(−13.2)
	G_4	−1.9	−3.0	−0.2	−11.9	−5.8	−3.9	−9.1	−11.2	−11.9
		(10.0)	(6.9)	(14.1)	(−7.0)	(−11.6)	(−9.5)	(−12.6)	(−16.1)	(−18.2)
	G_5	−3.4	−6.4	−2.4	−7.9	−5.4	−2.8	−8.5	−9.3	−9.8
		(6.4)	(−0.6)	(8.8)	(−7.3)	(−10.4)	(−9.7)	(−16.1)	(−18.4)	(−20.0)

survey data, we applied the PTFH model:

$$h_\lambda(y_i) = \beta_0 + \beta_1 \log x_i + v_i + \varepsilon_i, \quad i = 1, \ldots, 47,$$

where $v_i \sim N(0, A)$, $\varepsilon_i \sim N(0, D_i)$ and β_0, β_1, A and λ are model parameters. For comparison, we also applied the log-FH model corresponding $\lambda = 0$ in the above model, and the classical Fay-Herriot model. The model parameters were estimated by the maximum likelihood method in all models. For computing D_i in each model, we used the past data for consecutive eight years from 2006 to 2013, which are denoted by z_{it} for $t = 1, \ldots, 8$. In the FH and log-FH models, we simply calculated the sampling variance of $\{z_{i1}, \ldots, z_{i8}\}$ and $\{\log z_{i1}, \ldots, \log z_{i8}\}$,

Table 2.4: The CV values of the MSE estimator with known D_i and unknown D_i (Parenthesis) in each group.

		Pattern (a)			Pattern (b)			Pattern (c)		
	λ	0.2	0.6	1.0	0.2	0.6	1.0	0.2	0.6	1.0
max	G_1	1.15	1.03	1.14	0.78	0.60	0.66	0.77	0.62	0.67
		(2.72)	(1.65)	(1.96)	(2.13)	(1.78)	(1.89)	(2.01)	(1.19)	(1.29)
	G_2	1.17	1.33	1.50	0.55	0.73	0.71	0.54	0.50	0.54
		(2.25)	(2.50)	(3.42)	(1.20)	(2.44)	(2.49)	(0.91)	(0.84)	(0.82)
	G_3	2.25	0.98	1.16	0.59	0.87	1.00	0.52	0.46	0.53
		(1.34)	(1.80)	(2.16)	(0.88)	(1.02)	(1.1)	(0.68)	(0.98)	(0.97)
	G_4	1.02	1.26	1.54	0.46	0.84	0.98	0.51	0.50	0.57
		(1.44)	(1.40)	(2.08)	(0.70)	(0.94)	(1.06)	(0.66)	(1.72)	(1.70)
	G_5	0.72	1.01	1.23	0.67	0.73	0.88	0.47	0.66	0.77
		(1.10)	(1.59)	(2.21)	(0.91)	(1.60)	(1.87)	(0.56)	(0.90)	(0.91)
mean	G_1	0.99	0.74	0.82	0.51	0.50	0.55	0.45	0.41	0.43
		(1.05)	(1.33)	(1.57)	(1.07)	(1.07)	(1.15)	(1.10)	(0.86)	(0.90)
	G_2	0.78	0.91	1.06	0.46	0.56	0.63	0.40	0.42	0.47
		(1.23)	(1.46)	(1.83)	(0.80)	(1.05)	(1.11)	(0.73)	(0.66)	(0.68)
	G_3	1.02	0.86	1.03	0.46	0.61	0.71	0.42	0.41	0.45
		(1.08)	(1.27)	(1.60)	(0.76)	(0.81)	(0.88)	(0.60)	(0.70)	(0.72)
	G_4	0.75	0.93	1.13	0.43	0.66	0.78	0.41	0.44	0.50
		(1.10)	(1.24)	(1.65)	(0.64)	(0.80)	(0.89)	(0.56)	(0.85)	(0.87)
	G_5	0.71	0.92	1.12	0.52	0.66	0.77	0.41	0.49	0.56
		(0.99)	(1.36)	(1.83)	(0.70)	(0.94)	(1.05)	(0.53)	(0.69)	(0.72)
min	G_1	0.61	0.58	0.65	0.37	0.41	0.45	0.35	0.35	0.36
		(0.95)	(0.93)	(1.12)	(0.69)	(0.86)	(0.91)	(0.71)	(0.73)	(0.76)
	G_2	0.57	0.73	0.86	0.39	0.49	0.56	0.34	0.39	0.41
		(0.89)	(1.10)	(1.28)	(0.64)	(0.75)	(0.80)	(0.57)	(0.58)	(0.59)
	G_3	0.63	0.79	0.88	0.40	0.52	0.60	0.37	0.39	0.43
		(0.95)	(1.11)	(1.35)	(0.61)	(0.72)	(0.80)	(0.54)	(0.55)	(0.58)
	G_4	0.66	0.81	0.99	0.41	0.59	0.69	0.37	0.42	0.45
		(0.97)	(1.07)	(1.33)	(0.61)	(0.69)	(0.76)	(0.49)	(0.55)	(0.59)
	G_5	0.68	0.80	1.00	0.43	0.58	0.69	0.38	0.42	0.47
		(0.92)	(1.06)	(1.40)	(0.59)	(0.74)	(0.83)	(0.50)	(0.55)	(0.59)

respectively. In the PTFH model, similarly to Section 5.4.1, we first maximize (2.5) with $D_i = D_i(\lambda)$ and let $D_i = D_i(\widehat{\lambda})$.

In the PTFH model, we have $\widehat{\lambda} = 0.59$ in "Education" and $\widehat{\lambda} = 0.86$ in "Health". Moreover, based on based on 1000 parametric bootstrap samples, we obtained 95% confidence intervals of λ, $(0.20, 1.16)$ in "Education" and $(0.18, 1.99)$ in "Health", which indicate the log-transformation might not be appropriate. In Figure 2.1, we present the estimated regression lines of the three models, noting that $y = h_{\widehat{\lambda}}^{-1}(\widehat{\beta}_0 + \widehat{\beta}_1 x)$ in PTFH, and $y = \exp(\widehat{\beta}_0 + \widehat{\beta}_1 x)$ in log-FH. From the figure, it is observed that all the regression lines are similar. For assessing the suitability of normality assumptions of error terms, we computed the standardized resid-

uals: $e_i = r_i/\sqrt{\hat{A} + D_i}$, where r_i is the estimates of $v_i + \varepsilon_i$, so that $r_i = h_{\widehat{\lambda}}(y_i) - \widehat{\beta}_0 - \widehat{\beta}_1 x_i$ in PTFH, $r_i = \log y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i$ in log-FH and $r_i = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i$ in FH, noting that e_i asymptotically follows the standard normal distribution if the model specification is correct. In Figure 2.2, we give the estimated density of e_i in each model, which does not strongly supports the normality assumption of three models, but all the estimated densities are close to symmetric. Hence, the normality assumption might be plausible. In fact, we calculated the p-value of the Kolmogorov-Smirnov test for normality of e_i, presented in Table 2.5, and found that the normality assumption was not rejected in the three models in both items. Moreover, in Table 2.5, we provide AICs based on the maximum marginal likelihood for the three models. It can be seen that AICs of PTFH and log-FH are similar and smaller that that of FH in "Education" while that of PTFH is the smallest in "Health".

For investigation of goodness-of-fit of the PTFH model, we set $z_i = h_{\widehat{\lambda}}(y_i)$ and $w_i = \log x_i$, and applied the penalized spline model used in Opsomer et al. (2008):

$$z_i = \beta_0 + \sum_{\ell=1}^{p} \beta_\ell w_i^\ell + \sum_{\ell=1}^{K} \gamma_\ell (w_i - \kappa_\ell)_+^p + v_i + \varepsilon_i, \quad i = 1, \ldots, m, \tag{2.11}$$

where $v_i \sim (0, A)$, $\varepsilon_i \sim N(0, D_i)$ and $(\gamma_1, \ldots, \gamma_K)^t \sim N(0, \alpha \boldsymbol{I}_K)$, $(x)_+^p$ denotes the function $x^p I(x > 0)$, and $\kappa_1 < \cdots < \kappa_K$ is a set of fixed knots which determine the flexibility of splines. We set $K = 20$ and took κ_1 and κ_K as 10% and 90% quantiles of w_i, respectively, and set $\kappa_2, \ldots, \kappa_{K-1}$ at regular interval. For the degree of splines, we considered three cases: $p = 1, 2, 3$. We estimated model parameters $\beta_0, \ldots, \beta_p, A$ and α by the maximum likelihood method. In Figure 2.3, we present the estimated regression lines of three penalized spline models ($p = 1, 2, 3$) as well as that of PTFH, which shows that the linear parametric structure in the PTFH model seems plausible and PTFH would fit well in both items.

Finally, we computed the MSE estimates of the small area estimators for the three models. In the PTFH model, we used the estimator given in Theorem 2.1 with 1000 bootstrap samples and 5000 Monte Carlo samples of v_i and ε_i for numerical evaluation of g_{1i}. For the MSE estimates in the log-FH and FH models, we used the estimator given in Slud and Maiti (2006) and Datta and Lahiri (2000), respectively. We report the small area estimates and MSE estimates in seven prefectures around Tokyo in Table 2.6. It can be seen that log-FH and FH produce relatively similar estimates of area means while the estimates from PTFH are not similar to those models. Regarding MSE estimates, we can observe that the values in PTFH are smaller than the other two models, but we cannot directly compare these results since each MSE estimates are calculated based on the different sampling variances D_i.

Table 2.5: AIC and p-value of Kolmogorov-Smirnov (KS) test for normality of standardized residuals.

	AIC			p-value of KS test		
Data	PTFH	log-FH	FH	PTFH	log-FH	FH
Education	313.1	312.9	314.5	0.577	0.469	0.848
Health	172.9	180.7	183.4	0.519	0.440	0.375

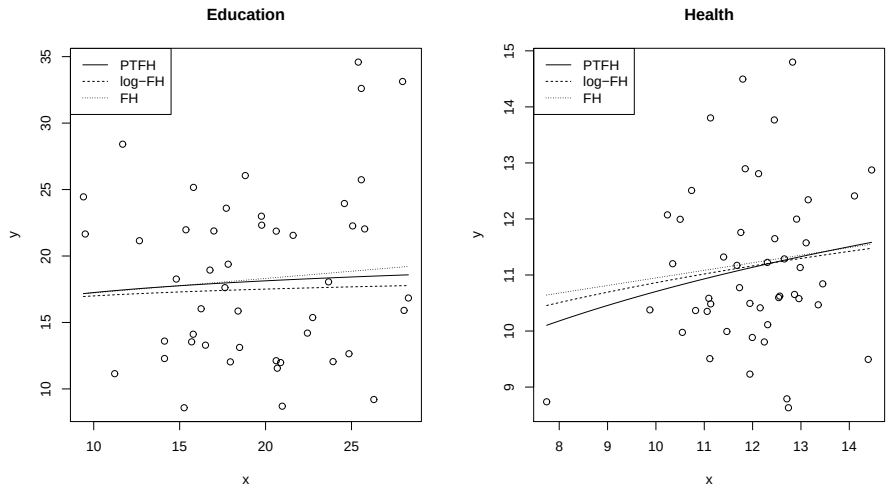

Figure 2.1: The scatter plots of (x_i, y_i) with estimated regression lines in the parametric transformed Fay-Herriot (PTFH) model, the log-transformed Fay-Herriot (log-FH) model and the classical Fay-Herriot (FH) model.

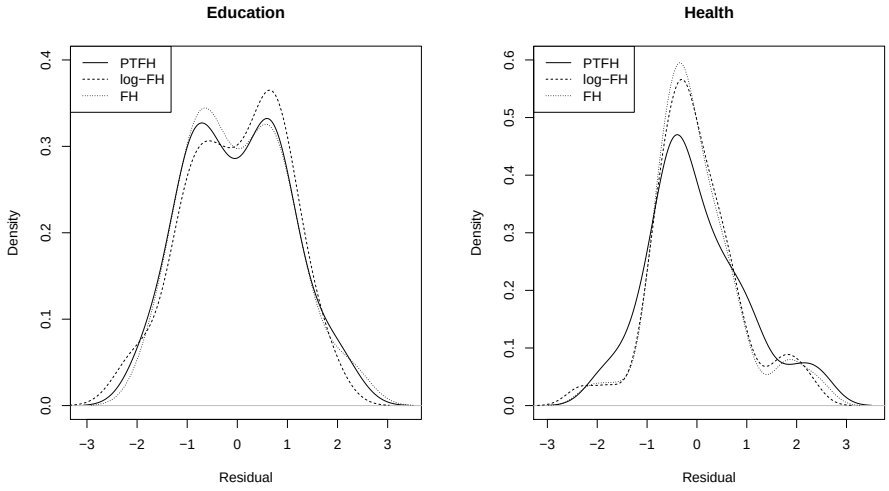

Figure 2.2: The estimated density of standardized residuals in the parametric transformed Fay-Herriot (PTFH) model, the log-transformed Fay-Herriot (log-FH) model and the classical Fay-Herriot (FH) model.

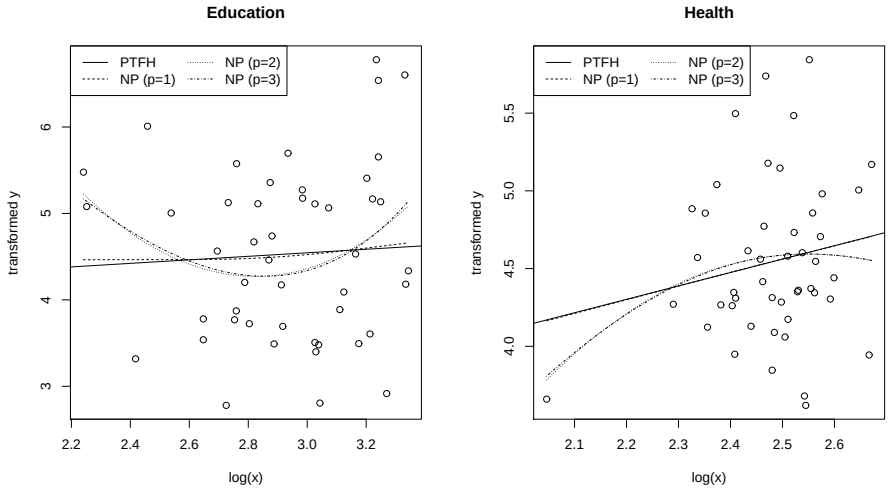

Figure 2.3: The scatter plots of $(\log x_i, h_{\widehat{\lambda}}(y_i))$ with estimated regression lines in the parametric transformed Fay-Herriot (PTFH) model and the nonparametric (NP) model based on the penalized spline with three orders $(p = 1, 2, 3)$.

Table 2.6: The small area estimates and the root of MSE (RMSE) estimates in seven prefectures around Tokyo from four models, the parametric transformed Fay-Herriot (PTFH) model, the log-transformed Fay-Herriot (log-FH) model and the classical Fay-Herriot (FH) model.

			Estimates			RMSE		
Data	Prefecture	DE	PTFH	log-FH	FH	PTFH	log-FH	FH
Education	Ibaraki	21.97	21.80	21.54	21.44	1.12	1.99	1.95
	Tochigi	21.88	21.63	21.21	21.30	1.65	2.41	2.10
	Gunma	14.12	14.76	15.49	15.17	2.74	3.68	2.93
	Saitama	32.61	27.41	23.81	22.78	4.49	4.68	4.83
	Chiba	21.55	20.92	20.19	20.08	3.31	3.90	3.87
	Tokyo	22.04	21.84	21.55	21.06	1.73	2.16	3.12
	Kanagawa	22.32	21.87	21.32	20.86	2.37	2.93	3.34
Health	Ibaraki	10.35	10.37	10.67	10.70	0.25	0.83	0.85
	Tochigi	11.76	11.71	11.34	11.33	0.61	1.08	1.08
	Gunma	8.74	8.88	10.00	10.12	0.51	1.23	1.25
	Saitama	11.13	11.13	11.21	11.22	0.19	0.77	0.79
	Chiba	12.81	12.64	11.64	11.64	0.60	1.06	1.07
	Tokyo	13.80	13.73	12.66	12.53	0.18	0.79	0.85
	Kanagawa	14.50	14.45	13.55	13.61	0.10	0.63	0.62

2.5 Technical Issues

2.5.1 Derivation of (2.6)

Note that $\mathrm{E}[(\widetilde{\mu}_i - \mu_i)^2] = \mathrm{E}[\mu_i^2] - \mathrm{E}[\widetilde{\mu}_i^2]$. From (2.4), it follows that

$$\mathrm{E}[\widetilde{\mu}_i^2] = \iiint_{\mathbb{R}^3} h_\lambda^{-1}(s) h_\lambda^{-1}(t) \phi(s; \widetilde{\theta}_i(u), \sigma_i^2) \phi(t; \widetilde{\theta}_i(u), \sigma_i^2) \phi(u; \boldsymbol{x}_i^t \boldsymbol{\beta}, A + D_i) ds dt du,$$

where $\widetilde{\theta}_i(u) = a_i u + (1-a_i)\boldsymbol{x}_i^t\boldsymbol{\beta}$ with $a_i = A/(A+D_i)$. Let S and T be random variables mutually independently distributed as $N(\widetilde{\theta}_i(U), \sigma_i^2)$ under given $U = u$, and let U be a random variable distributed as $N(\boldsymbol{x}_i^t\boldsymbol{\beta}, A + D_i)$. The marginal distribution of the vector $(S,T)^t$ is

$$N_2\left(\begin{pmatrix} \boldsymbol{x}_i^t\boldsymbol{\beta} \\ \boldsymbol{x}_i^t\boldsymbol{\beta} \end{pmatrix}, A\begin{pmatrix} 1 & a_i \\ a_i & 1 \end{pmatrix}\right).$$

Then, we have $\mathrm{E}[\widetilde{\mu}_i^2] = \mathrm{E}[h_\lambda^{-1}(S)h_\lambda^{-1}(T)]$, where the expectation is taken with respect to the marginal distribution of $(S,T)^t$. Introducing random variables z_1 and z_2 mutually independently distributed as $N(0,A)$, we can express $S = \boldsymbol{x}_i^t\boldsymbol{\beta} + c_{1i}z_1 + c_{2i}z_2$ and $T = \boldsymbol{x}_i^t\boldsymbol{\beta} + c_{1i}z_1 - c_{2i}z_2$, thereby we obtain the expression

$$\mathrm{E}[\widetilde{\mu}_i^2] = \mathrm{E}[h_\lambda^{-1}(\boldsymbol{x}_i^t\boldsymbol{\beta} + c_{1i}z_1 + c_{2i}z_2)h_\lambda^{-1}(\boldsymbol{x}_i^t\boldsymbol{\beta} + c_{1i}z_1 - c_{2i}z_2)].$$

Since $\mathrm{E}[\mu_i^2]$ can be expressed as $\mathrm{E}[\mu_i^2] = \mathrm{E}[\{h_\lambda^{-1}(\boldsymbol{x}_i^t\boldsymbol{\beta} + z_1)\}^2]$, we obtain (2.6).

2.5.2 Proof of Lemma 2.2

For notational simplicity, we define $\widetilde{\mu}_{i(\boldsymbol{\phi})} = \partial\widetilde{\mu}_i/\partial\boldsymbol{\phi}$ and $\widetilde{\mu}_{i(\boldsymbol{\phi}\boldsymbol{\phi})} = \partial^2\widetilde{\mu}_i/\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}^t$. Expanding $\widehat{\mu}_i$ around $\widetilde{\mu}_i$, we get

$$\widehat{\mu}_i - \widetilde{\mu}_i = \widetilde{\mu}_{i(\boldsymbol{\phi})}^t(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}) + \frac{1}{2}(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi})^t\widetilde{\mu}_{i(\boldsymbol{\phi}\boldsymbol{\phi})}(y_i;\boldsymbol{\phi}^*)(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}),$$

where $\boldsymbol{\phi}^*$ is on the line connecting $\boldsymbol{\phi}$ and $\widehat{\boldsymbol{\phi}}$. Then, it holds that

$$g_{2i}(\boldsymbol{\phi}) = \mathrm{E}\left[(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi})^t\widetilde{\mu}_{i(\boldsymbol{\phi})}\widetilde{\mu}_{i(\boldsymbol{\phi})}^t(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi})\right] + R_1 + \frac{1}{4}R_2,$$

where $R_1 = \mathrm{E}[\widetilde{\mu}_{i(\boldsymbol{\phi})}^t(\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi})(\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi})^t\widetilde{\mu}_{i(\boldsymbol{\phi}\boldsymbol{\phi})}(y_i;\boldsymbol{\phi}^*)(\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi})]$ and $R_2 = \mathrm{E}[\{(\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi})^t\widetilde{\mu}_{i(\boldsymbol{\phi}\boldsymbol{\phi})}(y_i;\boldsymbol{\phi}^*)(\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi})\}]$. We first show that $R_1 = o(m^{-1})$ and $R_2 = o(m^{-1})$. We only prove $R_1 = o(m^{-1})$ since the evaluation of R_2 is quite similar. In what follows, we define $\partial^2\widetilde{\mu}_i^*/\partial\phi_k\partial\phi_\ell = \partial^2\widetilde{\mu}_i(y_i;\boldsymbol{\phi}^*)/\partial\phi_k\partial\phi_\ell$. It follows that

$$\begin{aligned} R_1 &= \sum_{j=1}^{p+2}\sum_{k=1}^{p+2}\sum_{\ell=1}^{p+2} E\left[\left(\frac{\partial\widetilde{\mu}_i}{\partial\phi_j}\right)\left(\frac{\partial^2\widetilde{\mu}_i^*}{\partial\phi_k\partial\phi_\ell}\right)(\widehat{\phi}_j - \phi_j)(\widehat{\phi}_k - \phi_k)(\widehat{\phi}_\ell - \phi_\ell)\right] \\ &\equiv \sum_{j=1}^{p+2}\sum_{k=1}^{p+2}\sum_{\ell=1}^{p+2} U_{1jk\ell}, \end{aligned}$$

and

$$\begin{aligned} |U_{1jkl}| &\le \mathrm{E}\left[\left|\left(\frac{\partial\widetilde{\mu}_i}{\partial\phi_j}\right)\left(\frac{\partial^2\widetilde{\mu}_i^*}{\partial\phi_k\partial\phi_\ell}\right)\right|^4\right]^{\frac{1}{4}} \mathrm{E}\left[\left|(\widehat{\phi}_j - \phi_j)(\widehat{\phi}_k - \phi_k)(\widehat{\phi}_\ell - \phi_\ell)\right|^{\frac{4}{3}}\right]^{\frac{3}{4}} \\ &\le \mathrm{E}\left[\left|\frac{\partial\widetilde{\mu}_i}{\partial\phi_j}\right|^8\right]^{\frac{1}{8}} \mathrm{E}\left[\left|\frac{\partial^2\widetilde{\mu}_i^*}{\partial\phi_k\partial\phi_\ell}\right|^8\right]^{\frac{1}{8}} \prod_{a\in\{j,k,\ell\}} \mathrm{E}\left[\left|\widehat{\phi}_a - \phi_a\right|^4\right]^{\frac{1}{4}} \end{aligned}$$

from Hölder's inequality. From the asymptotic normality of $\widehat{\boldsymbol{\phi}}$ given in Lemma 8.1, it follows $\mathrm{E}[|\widehat{\phi} - \phi|^r] = O(m^{-r/2})$ for arbitrary $r > 0$. Then, we have

$$\prod_{a\in\{j,k,\ell\}} \mathrm{E}\left[\left|\widehat{\phi}_a - \phi_a\right|^4\right]^{1/4} = o(m^{-1}).$$

Noting that

$$\begin{aligned}\frac{\partial h_\lambda^{-1}(x)}{\partial\lambda} &\equiv \frac{\partial}{\partial\lambda}\left(\lambda x + \sqrt{1+\lambda^2x^2}\right)^{1/\lambda} \\ &= \frac{h_\lambda^{-1}(x)}{\lambda}\left\{\frac{x}{\sqrt{1+\lambda^2x^2}} - \frac{1}{\lambda}\log(\lambda x + \sqrt{1+\lambda^2x^2})\right\},\end{aligned}$$

the straightforward calculation shows that

$$\begin{aligned}\widetilde{\mu}_{i(\lambda)} &= \int_{-\infty}^{\infty}\left(\frac{\partial h_\lambda^{-1}(t)}{\partial\lambda}\right)\phi(t;\widetilde{\theta}_i,\sigma_i^2)dt + \int_{-\infty}^{\infty} h_\lambda^{-1}(t)\left(\frac{\partial\phi(t;\widetilde{\theta}_i,\sigma_i^2)}{\partial\lambda}\right)dt \\ &= \frac{1}{\lambda}\mathrm{E}\left[\frac{\theta_i h_\lambda^{-1}(\theta_i)}{\sqrt{1+\lambda^2\theta_i^2}}\Big|y_i\right] - \frac{1}{\lambda^2}\mathrm{E}\left[h_\lambda^{-1}(\theta_i)\log(\lambda\theta_i + \sqrt{1+\lambda^2\theta_i^2})\Big|y_i\right] \\ &\quad + \frac{1}{D_i}\left(\frac{\partial h_\lambda(y_i)}{\partial\lambda}\right)\mathrm{E}\left[(\theta_i - \widetilde{\theta}_i)h_\lambda^{-1}(\theta_i)\Big|y_i\right] \\ &\equiv \mathrm{E}[f_1(\theta_i)|y_i] + \mathrm{E}[f_2(\theta_i)|y_i] + \mathrm{E}[f_3(\theta_i,y_i)|y_i],\end{aligned}$$

where

$$\frac{\partial h_\lambda(y_i)}{\partial\lambda} = \frac{\log x}{\lambda}x^\lambda - \left(\log x + \frac{1}{\lambda}\right)h_\lambda(x).$$

Note that

$$\mathrm{E}\left[\left\{\mathrm{E}[f(\theta_i,y_i)|y_i]\right\}^a\right] \le \mathrm{E}\left[\mathrm{E}[f(\theta_i,y_i)^a|y_i]\right] = \mathrm{E}[f(\theta_i,y_i)^a]$$

for $a > 0$ from Jensen's inequality. Since $\mathrm{E}[f_1(\theta_i)^a] < \infty$, $\mathrm{E}[f_2(\theta_i)^a] < \infty$, $\mathrm{E}[f_3(\theta_i,y_i)^a] < \infty$ for $a > 0$, it follows that

$$\mathrm{E}\left[\left|\frac{\partial\widetilde{\mu}_i}{\partial\lambda}\right|^8\right] < \infty.$$

Similarly, we have

$$\begin{aligned}\widetilde{\mu}_{i(\beta)} &= \frac{D_i\boldsymbol{x}_i}{(A+D_i)\sigma_i^2}\mathrm{E}\left[(\theta_i - \widetilde{\theta}_i)h_\lambda^{-1}(\theta_i)\Big|y_i\right] \\ \widetilde{\mu}_{i(A)} &= \frac{D_i^2}{2\sigma_i^4(A+D_i)^2}\mathrm{E}\left[\left\{(\theta_i - \widetilde{\theta}_i)^2 - \sigma_i^{5/2}\right\}h_\lambda^{-1}(\theta_i)\Big|y_i\right],\end{aligned}$$

which leads to

$$\mathrm{E}\left[\left|\frac{\partial\widetilde{\mu}_i}{\partial\phi_k}\right|^8\right] < \infty,$$

for $k = 1, \ldots, p+1$. Moreover, straightforward but considerable calculations shows that $\mathrm{E}\left[|\partial^2 \widetilde{\mu}_i^* / \partial\phi_k \partial\phi_\ell|^8\right] < \infty$. Hence, we have $R_1 = o(m^{-1})$. A quite similar evaluation shows that $R_2 = o(m^{-1})$, which leads to

$$g_{2i}(\boldsymbol{\phi}) = \mathrm{E}\left[(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi})^t \widetilde{\mu}_{i(\boldsymbol{\phi})} \widetilde{\mu}_{i(\boldsymbol{\phi})}^t (\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi})\right] + o(m^{-1}).$$

Finally, using the similar argument given in the proof of Theorem 3 in Kubokawa et al. (2016), we have

$$\begin{aligned}\mathrm{E}\left[(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi})^t \widetilde{\mu}_{i(\boldsymbol{\phi})} \widetilde{\mu}_{i(\boldsymbol{\phi})}^t (\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi})\right] &= \mathrm{tr}\left(\mathrm{E}\left[\widetilde{\mu}_{i(\boldsymbol{\phi})} \widetilde{\mu}_{i(\boldsymbol{\phi})}^t\right] \mathrm{E}[(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi})(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi})^t]\right) + o(m^{-1}) \\ &= \frac{1}{m}\mathrm{tr}\left\{\boldsymbol{V}(\boldsymbol{\phi})\mathrm{E}\left[\widetilde{\mu}_{i(\boldsymbol{\phi})} \widetilde{\mu}_{i(\boldsymbol{\phi})}^t\right]\right\} + o(m^{-1}),\end{aligned}$$

which completes the proof.

2.5.3 Proof of Theorem 2.1

Taylor series expansion of $g_{1i}(\widehat{\boldsymbol{\phi}})$ around $\boldsymbol{\phi}$ gives

$$\mathrm{E}[g_{1i}(\widehat{\boldsymbol{\phi}})] = g_{1i}(\boldsymbol{\phi}) + \frac{\partial g_{1i}(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}^t}\mathrm{E}[\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}] + \frac{1}{2}\mathrm{tr}\left(\frac{\partial^2 g_{1i}(\boldsymbol{\phi})}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}^t}\mathrm{E}[(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi})(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi})^t]\right) + R_3,$$

where

$$R_3 = \frac{1}{6}\sum_{j=1}^{p+2}\sum_{k=1}^{p+2}\sum_{\ell=1}^{p+2} \frac{\partial^3 g_{1i}(\boldsymbol{\phi})}{\partial\phi_j\partial\phi_k\partial\phi_\ell}\bigg|_{\boldsymbol{\phi}=\boldsymbol{\phi}_*} (\widehat{\phi}_j - \phi_j)(\widehat{\phi}_k - \phi_k)(\widehat{\phi}_\ell - \phi_\ell).$$

Since $\mathrm{E}[(\widehat{\phi}_j - \phi_j)(\widehat{\phi}_k - \phi_k)(\widehat{\phi}_\ell - \phi_\ell)] = o(m^{-1})$, it holds that $\mathrm{E}[R_3] = o(m^{-1})$. Moreover, from Lemma 8.1, we have

$$\mathrm{E}[g_{1i}(\widehat{\boldsymbol{\phi}}) - g_{1i}(\boldsymbol{\phi})] = \frac{1}{m}\frac{\partial g_{1i}(\boldsymbol{\phi})}{\partial\boldsymbol{\phi}^t}\boldsymbol{b}(\boldsymbol{\phi}) + \frac{1}{2m}\mathrm{tr}\left(\frac{\partial^2 g_{1i}(\boldsymbol{\phi})}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}^t}\boldsymbol{V}(\boldsymbol{\phi})\right) + o(m^{-1}),$$

thereby, we have $\mathrm{E}[g_{1i}(\widehat{\boldsymbol{\phi}}) - g_{1i}(\boldsymbol{\phi})] = m^{-1}c_2(\boldsymbol{\phi}) + o(m^{-1})$ with the smooth function $c_2(\boldsymbol{\phi})$. Hence, from Lemma 2.2 and Butar and Lahiri (2003), we obtain the second order unbiasedness of (2.8).

Chapter 3

Adaptively Transformed Mixed Model Prediction

3.1 Introduction

We consider a finite population partitioned m areas and each area has N_i populations for $i = 1, \ldots, m$. Let Y_{ij} be the characteristics of jth individuals in ith area. We are interested in the area mean:

$$\mu_i = \frac{1}{N_i} \sum_{j=1}^{N_i} T(Y_{ij}), \tag{3.1}$$

where T is a known (user-specified) function. For example, in poverty mapping, we often use $T_\alpha(x) = \{(z-x)/z\}^\alpha I(x < z)$ known as FGT poverty measure (Foster et al., 1984). In this case, μ_i represents poverty rate ($\alpha = 0$), poverty gap ($\alpha = 1$) and poverty severity ($\alpha = 2$) in ith area. If we could observed all the units Y_{ij}, we could calculate the true value of μ_i. However, in practice, we can only observe $n_i (< N_i)$ units in each area. Since the sample size n_i is small compared with N_i, the direct estimator of μ_i based only on the sampled data:

$$\widehat{\mu}_i^D = \frac{1}{n_i} \sum_{j=1}^{n_i} T(y_{ij})$$

has a large variance and produces an inaccurate estimate. In most applications, some covariates $\boldsymbol{x}_{ij}$ associated with Y_{ij} are available for sampled as well as non-sampled units. Under the setting, Molina and Rao (2010) proposed an empirical best prediction (EBP) method for μ_i using the nested error regression model:

$$Y_{ij} = \boldsymbol{x}_{ij}^t \boldsymbol{\beta} + v_i + \varepsilon_{ij}, \quad j = 1, \ldots, N_i, \quad i = 1, \ldots, m, \tag{3.2}$$

where $\boldsymbol{x}_{ij}$ and $\boldsymbol{\beta}$ are p-dimensional vectors of covariates and regression coefficients, v_i is the area-specific effect which follows $N(0, \tau^2)$ and ε_{ij} is a sampling error distributed as $N(0, \sigma^2)$. The model parameters are estimated from the sampled data $\boldsymbol{y}_s = \{y_{ij},\ j = 1, \ldots, n_i,\ i = 1, \ldots, m\}$. Under the model (3.2), the conditional distribution of $Y_{ij},\ j = n_i + 1, \ldots, N_i$ given $\boldsymbol{y}_s$ is

$$Y_{ij} | \boldsymbol{y}_s \sim N\left(\boldsymbol{x}_{ij}^t \boldsymbol{\beta} + \frac{n_i \tau^2}{\sigma^2 + n_i \tau^2} (\bar{y}_i - \bar{\boldsymbol{x}}_i^t \boldsymbol{\beta}),\ \frac{\sigma^2 \tau^2}{\sigma^2 + n_i \tau^2} \right), \quad j = n_i + 1, \ldots, N_i. \tag{3.3}$$

Then the best predictor of μ_i can be obtained as the conditional expectation $\mathrm{E}[\mu_i|y_i]$ which has the form

$$\widetilde{\mu}_i = \mathrm{E}[\mu_i|\boldsymbol{y}_i] = \frac{1}{N_i}\left\{\sum_{j=1}^{n_i} T(y_{ij}) + \sum_{j=n_i+1}^{N_i} \mathrm{E}[T(Y_{ij})|\boldsymbol{y}_i]\right\}.$$

Though the expectation $\mathrm{E}[T(Y_{ij})|\boldsymbol{y}_s]$ cannot be expressed in a closed form for general function $T(\cdot)$, it can be easily computed via Monte Calro integration by generating large number of random samples from the conditional distribution (3.3). The empirical best predictor $\widehat{\mu}_i$ is obtained by replacing unknown model parameters in $\widetilde{\mu}_i$ with some estimator. Molina and Rao (2010) demonstrated that the empirical best predictor performs quite well compared with the direct estimator as well as ELL method (Elbers et al., 2003), the standard method for poverty mapping used in World Bank.

The key assumption of the EBP method is the normality of the unit sample y_{ij}, which enables us to obtain the simple expression of the conditional distribution (3.3). However, when y_i is positive valued and its distribution is far from normality, the EBP method could be inefficient and biased. In Molina and Rao (2010), transformed variable $H(y_{ij})$ with some known function $H(\cdot)$ instead of y_{ij} is used in the nested error model (3.2). The selection of the transformation $H(\cdot)$ is an important issue since the misspecification of $H(\cdot)$ leads to inconsistency of the EBP method. To overcome the difficulty, in this chapter, we use the parametric family of transformations for y_{ij} and estimate the transformation parameter as well as the model parameters in (3.2) from the sampled data. In Section 3.2, we propose the nested error regression model with parametrically transformed response values, and an estimating method for the model parameters. Then we suggest the flexibly transformed empirical best predictor (FTEBP) of μ_i. For measuring the variability of FTEBP, we propose an empirical Bayes confidence interval of μ_i. In Section 3.4, we present some simulation studies and an example of the proposed method.

3.2 Adaptively Transformed Mixed Model Prediction

3.2.1 Transformed best predictor

Let $H_\lambda(\cdot)$ be a family of transformations with parameter λ. The transformation parameter λ might be multidimensional, but we treat λ as a scalar parameter for notational simplicity. The assumptions and specific choices of $H_\lambda(\cdot)$ will be discussed in the subsequent section. We assume that the transformed variable $H_\lambda(y_{ij})$ follows the nested error regression model:

$$H_\lambda(Y_{ij}) = \boldsymbol{x}_{ij}^t\boldsymbol{\beta} + v_i + \varepsilon_{ij}, \quad j = 1, \ldots, N_i, \quad i = 1, \ldots, m, \tag{3.4}$$

where $\boldsymbol{x}_{ij}$ and $\boldsymbol{\beta}$ are p-dimensional vectors of covariates and regression coefficients, v_i and ε_{ij} are an area-specific effect and a sampling error, respectively. Here we assume that v_i and ε_{ij} are mutually independent and distributed as $v_i \sim N(0, \tau^2)$ and $\varepsilon_{ij} \sim N(0, \sigma^2)$ with unknown two variance parameters τ^2 and σ^2. It is worth noting that, owing to the area effect v_i, the units in the same area are mutually correlated while the units in the different area are independent. Specifically, from (3.4), it holds $\mathrm{Cor}(H_\lambda(Y_{ij}), H_\lambda(Y_{ik})) = (\tau^2 + \sigma^2)^{-1}\tau^2,\ j \neq k$, thereby the units in the same area are mutually correlated and the degree of correlation is determined by the ratio τ^2/σ^2. From the normality assumptions of v_i and ε_{ij}, it follows

that $H_\lambda(Y_{ij}) \sim N(\boldsymbol{x}_{ij}^t\boldsymbol{\beta}, \tau^2+\sigma^2)$. Thus, the transformation parameter λ can be chosen to make the transformed data $H_\lambda(y_{ij})$ close to normality. We define $\boldsymbol{\phi} = (\boldsymbol{\beta}^t, \tau^2, \sigma^2, \lambda)^t$, as the vector of unknown model parameters in (3.4). The estimation procedure will be given in the subsequent section.

Let $y_s = \{y_{ij},\ j = 1, \ldots, n_i,\ i = 1, \ldots, m\}$ be the sampled data. From the model (3.4), we have $H_\lambda(Y_{ij})|y_s \sim N(\theta_{ij}, s_i^2+\sigma^2)$, $j = n_i+1, \ldots, N_i$, where

$$\theta_{ij} = \boldsymbol{x}_{ij}^t\boldsymbol{\beta} + \frac{\tau^2}{\sigma^2+n_i\tau^2}\sum_{j=1}^{n_i}(H_\lambda(y_{ij}) - \boldsymbol{x}_{ij}^t\boldsymbol{\beta}), \quad s_i = \sqrt{\frac{\sigma^2\tau^2}{\sigma^2+n_i\tau^2}}. \tag{3.5}$$

Hence, the best predictor of μ_i given in (3.1) can be obtained as

$$\widetilde{\mu}_i(y_s;\boldsymbol{\phi}) \equiv \mathrm{E}[\mu_i|y_s] = \frac{1}{N_i}\left\{\sum_{j=1}^{n_i} T(y_{ij}) + \sum_{j=n_i+1}^{N_i}\mathrm{E}[T\circ H_\lambda^{-1}(u_{ij})]\right\}, \tag{3.6}$$

where the expectation is taken with respect to $u_{ij} \sim N(\theta_{ij}, s_i^2+\sigma^2)$, and $T\circ H_\lambda^{-1}(\cdot)$ is the composite function of $T(\cdot)$ and H_λ^{-1}, the inverse function of $H_\lambda(\cdot)$. Although the expectation $\mathrm{E}[T\circ H_\lambda^{-1}(u_{ij})]$ does not have a closed form in general, it can be easily computed via the Monte Carlo integration. We call the best predictor (3.6) adaptively transformed best predictor (ATBP).

3.2.2 Estimation of structural parameters

We here consider estimating the unknown model parameters $\boldsymbol{\phi}$ in (3.4) based on the marginal likelihood function. Noting that the log-marginal likelihood function of $\boldsymbol{\phi}$ is given by

$$\begin{aligned} L(\boldsymbol{\phi}) = &-\frac{1}{2}\sum_{i=1}^m \log|\boldsymbol{\Sigma}_i| - \frac{1}{2}\sum_{i=1}^m \{H_\lambda(y_i) - \boldsymbol{X}_i\boldsymbol{\beta}\}^t\boldsymbol{\Sigma}_i^{-1}\{H_\lambda(y_i) - \boldsymbol{X}_i\boldsymbol{\beta}\} \\ &- \frac{1}{2}\sum_{i=1}^m n_i\log 2\pi + \sum_{i=1}^m\sum_{j=1}^{n_i}\log H_\lambda'(y_{ij}), \end{aligned} \tag{3.7}$$

where $(\boldsymbol{\Sigma}_i)_{k\ell} = \tau^2+\sigma^2 I(k=\ell)$, $H_\lambda(y_i) = (H_\lambda(y_{i1}), \ldots, H_\lambda(y_{in_i}))^t$, $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}^t, \ldots, \boldsymbol{x}_{in_i}^t)^t$, and $H_\lambda'(\cdot)$ denotes the derivative of $H_\lambda(\cdot)$. The maximum likelihood estimator of $\boldsymbol{\phi}$ can be defined as the maximizer of $L(\boldsymbol{\phi})$.

For maximizing the likelihood function $L(\boldsymbol{\phi})$, we first note that the profile likelihood function of λ can be expressed as

$$\mathrm{PL}(\lambda) = \mathrm{ML}(\lambda) + \sum_{i=1}^m\sum_{j=1}^{n_i}\log H_\lambda'(y_{ij}), \tag{3.8}$$

where $\mathrm{ML}(\lambda)$ is the maximum likelihood of the nested error regression model with response values $H_\lambda(y_{ij})$ and covariate vectors $\boldsymbol{x}_{ij}$, which can be efficiently carried out by using well-developed numerical method (e.g. Molina and Marhuenda, 2015). Using the ease of the point evaluation of the profile likelihood $\mathrm{PL}(\lambda)$, we can obtain the maximizer of $\mathrm{PL}(\lambda)$ by using, for example, the golden section method (Brent et al., 1973). Once we obtain the estimator $\widehat{\lambda}$,

we get the estimators of other parameters by applying the nested error regression model to the data set $\{H_{\widehat{\lambda}}(y_{ij}), \boldsymbol{x}_{ij}\}$.

For estimating the two variance parameters τ^2 and σ^2, the restricted maximum likelihood (RML) method (Jiang, 1996) might be more attractive than the maximum likelihood method. To implement the RML estimation, the first three terms in (3.7) need to be changed to the restricted maximum likelihood, but the transformation parameter λ can be easily estimated in the same manner as the maximum likelihood method based on the profile likelihood function. However, in this paper, we consider only the maximum likelihood estimator for simplicity.

3.2.3 Class of transformations

We here consider the concrete choice of the family of transformations $H_\lambda(\cdot)$. To begin with, we give some conditions to be satisfied by the transformations.

Assumption 3.1. *(Class of transformations)*

1. *H_λ is a differentiable and monotone function, and the range of H_λ is $\mathbb{R}$ for all λ.*
2. *For fixed x, $H_\lambda(x)$ as the function of λ is differentiable.*
3. *The function $|\partial H_\lambda(w)/\partial\lambda|$, $|\partial^2 H_\lambda(w)/\partial\lambda^2|$ and $|\partial^2 \log H'_\lambda(w)/\partial\lambda^2|$ with $w = H_\lambda^{-1}(x)$ are bounded from the upper by $C_1\{\exp(C_2 x)+\exp(-C_2 x)\}$ with some constants $C_1, C_2 > 0$.*

The first condition is crucial in this context. If the range of H_λ is not $\mathbb{R}$, but some subset $A \subset \mathbb{R}$, the inverse function H_λ^{-1} cannot be defined on $\mathbb{R} \setminus A$, which causes problems in computing the best predictor (3.6). When the observations are positive valued, the Box-Cox (BC) transformation (Box and Cox, 1964), $H_\lambda(x) = \lambda^{-1}(x^\lambda - 1)$ for $\lambda \neq 0$ and $H_0(x) = \log(x)$, is widely used. However, it is known that the range of BC transformation is truncated and not whole real line, so that the BC transformation cannot be used in this context. An alternative transformation, called dual power (DP) transformation, has been suggested by Yang (2006):

$$H_\lambda^{\mathrm{DP}}(x) = \frac{x^\lambda - x^{-\lambda}}{2\lambda}, \quad x > 0, \quad \lambda > 0, \tag{3.9}$$

where $\lim_{\lambda\to 0} H_\lambda^{\mathrm{DP}}(x) = \log x$. It can be seen as the mean of two BC transformations, and it is easy to confirm that the range of DPT is $\mathbb{R}$, so that DPT can be used as a parametric family including log-transformation in this context. The expression of the inverse function is required in computing the transformed best predictor (3.6), and the Jacobian is also needed for computing the profile likelihood function (3.8). These are given by

$$H_\lambda^{\mathrm{DP}(-1)}(x) = \left(\lambda x + \sqrt{1+\lambda^2 x^2}\right)^{1/\lambda} \quad \text{and} \quad \frac{dH_\lambda^{\mathrm{DP}}(x)}{dx} = \frac{1}{2}(x^{\lambda-1} + x^{-\lambda-1}).$$

In the context of small area estimation, the DP transformation was used in Sugasawa and Kubokawa (2017) in the Fay-Herriot model. The original DP transformation (3.9) can be used when the response variables are positive. When response variables are real valued, one

may use the shifted-DP transformation of the form $H_{\lambda,c}(x) = \{(x+c)^{\lambda} - (x+c)^{-\lambda}\}/2\lambda$, where $c \in (\min(y_{ij}) + \varepsilon, \infty)$ with specified small $\varepsilon > 0$.

Another attractive transformation is the sinh-arcsinh (SS) transformation suggested in Jones and Pewsey (2009) in the context of distribution theory, which has the form

$$H^{\mathrm{SS}}_{a,b}(x) = \sinh(b \sinh^{-1}(x) - a), \quad x \in (-\infty, \infty), \quad a \in (-\infty, \infty), \quad b \in (0, \infty) \tag{3.10}$$

where $\sinh(x) = (e^x - e^{-x})/2$ is the hyperbolic sine function, $\sinh^{-1}(x) = \log(x + \sqrt{x^2+1})$, and two transformation parameter a and b control skewness and tail heaviness, respectively. The inverse transformation and the Jacobian are obtained as

$$H^{\mathrm{SS}(-1)}_{a,b}(x) = \sinh(b^{-1} \sinh^{-1}(x) + a), \quad \text{and} \quad \frac{dH^{\mathrm{SS}}_{a,b}(x)}{dx} = b\sqrt{\frac{1 + H^{\mathrm{SS}}_{a,b}(x)^2}{1+x^2}}.$$

These transformations will be used and compared in the application presented in Section 3.4.3.

3.2.4 Large sample properties

We here consider the large sample properties of the estimator of structural parameters. To this end, we assume the following condition:

Assumption 3.2. *(Assumptions under large m)*

1. *The true parameter vector $\boldsymbol{\phi}_0$ is an interior point of the parameter space $\boldsymbol{\Phi}$.*
2. *$0 < \min_{i=1,\ldots,m} N_i \leq \max_{i=1,\ldots,m} N_i < \infty$.*
3. *The elements of $\boldsymbol{X}_i$ are uniformly bounded and $\boldsymbol{X}_i^t \boldsymbol{X}_i$ is positive definite.*
4. *$m^{-1} \sum_{i=1}^m \boldsymbol{X}_i^t \boldsymbol{\Sigma}_i^{-1} \boldsymbol{X}_i$ converges to a positive definite matrix as $m \to \infty$.*

Since the asymptotic variance and covariance matrix of MLE can be derived from the Fisher information matrix, we first provide the Fisher information matrix in the following Theorem, where the proof is given in Appendix.

Theorem 3.1. *We define the Fisher information $I_{\phi_k \phi_j} = -\mathrm{E}[\partial^2 L(\boldsymbol{\phi})/\partial\phi_k \partial\phi_j]$, then it follows that*

$$I_{\tau^2\tau^2} = \frac{1}{2}\sum_{i=1}^m (\mathbf{1}_{n_i}^t \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i})^2, \quad I_{\tau^2\sigma^2} = \frac{1}{2}\sum_{i=1}^m \mathbf{1}_{n_i}^t \boldsymbol{\Sigma}_i^{-2} \mathbf{1}_{n_i}, \quad I_{\sigma^2\sigma^2} = \frac{1}{2}\sum_{i=1}^m \mathrm{tr}(\boldsymbol{\Sigma}_i^{-2}),$$

$$\boldsymbol{I}_{\boldsymbol{\beta}\boldsymbol{\beta}} = \sum_{i=1}^m \boldsymbol{X}_i^t \boldsymbol{\Sigma}_i^{-1} \boldsymbol{X}_i, \quad \boldsymbol{I}_{\boldsymbol{\beta}\tau^2} = \boldsymbol{I}_{\boldsymbol{\beta}\sigma^2} = 0, \quad I_{\lambda\sigma^2} = -\sum_{i=1}^m \mathrm{E}\left[\boldsymbol{z}_i^t \boldsymbol{\Sigma}_i^{-2} H_\lambda^{(1)}(y_i)\right],$$

$$\boldsymbol{I}_{\lambda\boldsymbol{\beta}} = -\sum_{i=1}^m \boldsymbol{X}_i^t \boldsymbol{\Sigma}_i^{-1} \mathrm{E}\left[H_\lambda^{(1)}(y_i)\right], \quad I_{\lambda\tau^2} = -\sum_{i=1}^m \mathrm{E}\left[\boldsymbol{z}_i^t \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^t \boldsymbol{\Sigma}_i^{-1} H_\lambda^{(1)}(y_i)\right],$$

$$I_{\lambda\lambda} = \sum_{i=1}^m \mathrm{E}\left[H_\lambda^{(1)}(y_i)^t \boldsymbol{\Sigma}_i^{-1} H_\lambda^{(1)}(y_i)\right] + \sum_{i=1}^m \mathrm{E}\left[\boldsymbol{z}_i^t \boldsymbol{\Sigma}_i^{-1} H_\lambda^{(2)}(y_i)\right] - \sum_{i=1}^m \sum_{j=1}^{n_i} \mathrm{E}\left[\frac{\partial^2}{\partial\lambda^2} \log H'_\lambda(y_{ij})\right],$$

where $H_\lambda^{(k)}(y_i) = \partial^k H_\lambda(y_i)/\partial\lambda^k$ *for* $k = 1, 2$, $\boldsymbol{z}_i = H_\lambda(y_i) - \boldsymbol{X}_i\boldsymbol{\beta}$, *and* $\mathrm{E}[\cdot]$ *denotes the expectation with respect to* y_{ij}*'s following the model (3.4). Then, under Assumptions 3.1 and 3.2, the maximum likelihood estimator* $\widehat{\boldsymbol{\phi}}$ *is asymptotically distributed as* $\widehat{\boldsymbol{\phi}} \sim N(\boldsymbol{\phi}, \boldsymbol{I}_{\boldsymbol{\phi}}^{-1})$.

From Theorem 3.1, it is observed that the information matrix of $(\boldsymbol{\beta}^t, \tau^2, \sigma^2)$ does not depend on the transformation parameter λ, and their expressions are the same as those of the traditional nested error regression models. While the two variance parameters τ^2 and σ^2 are orthogonal to $\boldsymbol{\beta}$ in the sense that $\boldsymbol{I}_{\boldsymbol{\beta}\tau^2} = \boldsymbol{I}_{\boldsymbol{\beta}\sigma^2} = 0$, the transformation parameter λ is not orthogonal to the others. The expectations appeared in the Fisher matrix is not analytically tractable, but it can be easily estimated by replacing the expectation with its sample counterpart. In the case that λ is multidimensional, the extension of Theorem 3.1 is straightforward. The expressions of $H_\lambda^{(k)}(y_i)$ and $\partial^2 \log H'_\lambda(y_{ij})/\partial\lambda^2$ could be analytically complicated and require tedious algebraic calculations. In such a case, the numerical derivative can be useful since we need to compute only the point values of the derivatives.

3.3 Empirical Bayes confidence intervals

3.3.1 Asymptotically valid confidence intervals

Measuring the variability of the transformed empirical best predictor $\widehat{\mu}_i$ is an important issue in practice. Traditionally, the mean squared error (MSE) of $\widehat{\mu}_i$ has been used, and several methods ranging from analytical method (Prasad and Rao, 1990) to numerical methods (Hall and Maiti, 2006a) have been considered. On the other hand, an empirical Bayes confidence interval of μ_i is more preferable since it can provide distributional information than MSE though construction of the confidence interval is generally difficult. Here, we derive an asymptotically valid empirical Bayes confidence interval of μ_i.

The key to the confidence interval is the conditional distribution of μ_i given y_i. Noting that $\mathrm{Cov}(H_\lambda(Y_{ij}), H_\lambda(Y_{ik})|y_i) = \mathrm{Var}(v_i|y_i) = s_i^2$ for $j \neq k$, it follows that

$$(H_\lambda(Y_{i,n_i+1}), \ldots, H_\lambda(Y_{iN_i}))^t | y_i \sim N((\theta_{i,n_i+1}, \ldots, \theta_{iN_i})^t, s_i^2 \mathbf{1}_{N_i-n_i}\mathbf{1}_{N_i-n_i}^t + \sigma^2 \boldsymbol{I}_{N_i-n_i}),$$

namely, the each component has the expression

$$H_\lambda(Y_{ij})|y_i = \theta_{ij} + s_i z_i + \sigma w_{ij}, \quad j = n_i + 1, \ldots, N_i,$$

where z_i and w_{ij} are mutually independent standard normal random variables, and θ_{ij} and s_i are defined in (3.5). Then the posterior distribution of μ_i can be expressed as

$$\mu_i|y_i \stackrel{\mathrm{d}}{=} \frac{1}{N_i}\left\{\sum_{j=1}^{n_i} T(y_{ij}) + \sum_{j=n_i+1}^{N_i} T \circ H_\lambda^{-1}\left(\theta_{ij} + s_i z_i + \sigma w_{ij}\right)\right\}, \tag{3.11}$$

which is a complex function of standard normal random variables z_i and w_{ij}. However, random samples from the conditional distribution (3.11) can be easily simulated.

We define $Q_a(y_i, \boldsymbol{\phi})$ as the lower $100a\%$ quantile point of the posterior distribution of μ_i with the true $\boldsymbol{\phi}$, which satisfies $\mathrm{P}(\mu_i \leq Q_a(y_i, \boldsymbol{\phi})|y_i) = a$. Hence, the Bayes confidence interval of μ_i with nominal level $1 - \alpha$ is obtained as $I_\alpha = (Q_{\alpha/2}(y_i, \boldsymbol{\phi}), Q_{1-\alpha/2}(y_i, \boldsymbol{\phi}))$, which holds

that $\mathrm{P}(\mu_i \in I_\alpha) = 1 - \alpha$. However, the interval I_α depends on the unknown parameter $\boldsymbol{\phi}$, so that the feasible version of I_α is obtained by replacing $\boldsymbol{\phi}$ with its estimator $\widehat{\boldsymbol{\phi}}$, namely

$$I_\alpha^N = (Q_{\alpha/2}(y_i, \widehat{\boldsymbol{\phi}}), Q_{1-\alpha/2}(y_i, \widehat{\boldsymbol{\phi}})), \tag{3.12}$$

which we call naive empirical Bayes confidence interval of μ_i. The two quantiles appeared in (3.12) can be computed by generating a large number of random samples from the conditional distribution (3.11). Owing to the asymptotic properties of $\widehat{\boldsymbol{\phi}}$, the coverage probability of the naive interval (3.12) converges to the nominal level as the number of areas m tends to infinity as shown in the following theorem proved in Appendix.

Theorem 3.2. *Under Assumptions 3.1 and 3.2, it holds* $P(\mu_i \in I_\alpha^N) = 1 - \alpha + O(m^{-1})$.

3.3.2 Bootstrap calibrated intervals

As shown in Theorem 3.2, the coverage error of the naive interval (3.12) is of order m^{-1}, which is not necessarily negligible when m is not sufficiently large. Since the number of m is usually moderate in practice, the calibrated intervals with higher accuracy would be valuable. Following Chatterjee, et al. (2008), Hall and Maiti (2006a), we construct a second order corrected empirical Bayes confidence interval I_α^C satisfying $P(\mu_i \in I_\alpha^C) = 1 - \alpha + o(m^{-1})$.

To begin with, we define the bootstrap estimator of the coverage probability of the naive interval. Let Y_{ij}^* be the parametric bootstrap samples generated from the estimated model (3.4) with $\boldsymbol{\phi} = \widehat{\boldsymbol{\phi}}$, and $y_i^* = \{Y_{ij}^*,\ j = 1, \ldots, n_i\}$. Moreover, let μ_i^* be the bootstrap version of μ_i based on Y_{ij}^*'s. Since the coverage probability is $\mathrm{P}(Q_{a/2}(y_i, \widehat{\boldsymbol{\phi}}) \le \mu_i \le Q_{1-a/2}(y_i, \widehat{\boldsymbol{\phi}}))$, its parametric bootstrap estimator can be defined as

$$\mathrm{CP}(a) = \mathrm{E}^* \left[I \left\{ Q_{a/2}(y_i^*, \widehat{\boldsymbol{\phi}}^*) \le \mu_i^* \le Q_{1-a/2}(y_i^*, \widehat{\boldsymbol{\phi}}^*) \right\} \right],$$

where the expectation is taken with respect to the bootstrap samples Y_{ij}^*'s. Based on the coverage probability, we define the calibrated nominal level a^* as the solution of the equation $\mathrm{CP}(a^*) = 1 - \alpha$, which can be solved by the bisectional method (Brent, 1973). Then, the calibrated interval is given by

$$I_\alpha^C = (Q_{a^*/2}(y_i, \widehat{\boldsymbol{\phi}}), Q_{1-a^*/2}(y_i, \widehat{\boldsymbol{\phi}})), \tag{3.13}$$

which has second order accuracy as shown in the following theorem proved in Appendix.

Theorem 3.3. *Under Assumptions 3.1 and 3.2, it holds* $P(\mu_i \in I_\alpha^C) = 1 - \alpha + o(m^{-1})$.

3.4 Numerical Studies

3.4.1 Evaluation of prediction errors

We first evaluate the prediction errors of the proposed predictors together with some existing methods. To this end, we considered the following data generating processes:

$$\begin{aligned}
&\text{(A)}\quad (2\lambda)^{-1}(Y_{ij}^{\lambda}-Y_{ij}^{-\lambda})=\beta_0+\beta_1 X_{ij}+v_i+\varepsilon_{ij},\quad v_i\sim N(0,\tau^2),\quad \varepsilon_{ij}\sim N(0,\sigma^2)\\
&\text{(B)}\quad (2\lambda)^{-1}(Y_{ij}^{\lambda}-Y_{ij}^{-\lambda})=\beta_0+\beta_1 X_{ij}+v_i+\varepsilon_{ij},\quad v_i\sim t_5(0,\tau^2),\quad \varepsilon_{ij}\sim t_5(0,\sigma^2)\\
&\text{(C)}\quad Y_{ij}=\exp(\beta_0+\beta_1 X_{ij})v_i\varepsilon_{ij},\quad v_i\sim \Gamma(1/\tau^2,1/\tau^2),\quad \varepsilon_{ij}\sim \Gamma(1/\sigma^2,1/\sigma^2)\\
&\text{(D)}\quad Y_{ij}=0.2\exp(U_{ij})+0.8U_{ij}^2,\quad U_{ij}=\beta_0+\beta_1 X_{ij}+v_i+\varepsilon_{ij},\\
&\qquad\quad v_i\sim N(0,\tau^2),\quad \varepsilon_{ij}\sim N(0,\sigma^2),
\end{aligned}$$

where $i=1,\ldots,m$, $j=1,\ldots,N$, $\beta_0=-1$, $\beta_1=3$, $\tau=0.3$, $\sigma=0.7$, and X_{ij} were initially generated from $U(1,2)$ and fixed through simulation experiments. In model (i) and (ii), we considered three values for λ, $\lambda=0, 0.2$ and 0.4. In this study, we set $N=200$ and $m=25$, and we focus on estimating the ratio of the observation with values under z, namely

$$\mu_i=\frac{1}{N}\sum_{j=1}^{N} I(Y_{ij}<z),\qquad i=1,\ldots,m, \tag{3.14}$$

where z is defined as 0.6 times median of Y_{ij}'s.

Concerning the area sample sizes, we divided $m=25$ areas into five groups with equal number of areas, and we set the same number of n_i within the same groups. The group pattern of n_i we considered was $(20, 40, 60, 80, 100)$. Among the generated $Y_{i1},\ldots,Y_{iN}$, we used first n_i observations $y_{i1}(=Y_{i1}),\ldots,y_{in_i}(=Y_{in_i})$ as the sampled data. Then, based on the sampled data y_{ij}'s and covariates X_{ij}'s, we computed the predicted value of μ_i based on the four methods: the proposed flexible transformed empirical best prediction (ATP) method with DP transformation (3.9), the transformed empirical best prediction (TP) method proposed by Molina and Rao (2010) with log-transformation, the empirical best prediction (EBP) method by directly applying the nested error regression model to the non-transformed observation y_{ij}, and the direct estimator (DE) given by

$$\widehat{\mu}_i^{D}=\frac{1}{n_i}\sum_{j=1}^{n_i} I(y_{ij}<z),\qquad i=1,\ldots,m.$$

It should be noted that the TP method is correctly specified in scenario (A) with $\lambda=0$ while the ATP method is overfitting in this case. In the other cases in scenario (A), the ATP method uses the same model as the data generating model. Scenario (B) is similar to (A), but the distribution of error terms have the t-distribution. In scenario (C) and (D), the data generation models do not coincides with any methods.

To compare the performances of the four methods, we computed the square root of mean squared error (RMSE) defined as

$$\mathrm{RMSE}_i=\sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\widehat{\mu}_i^{(r)}-\mu_i^{(r)}\right)^2},$$

where $R = 2000$ in this study, $\widehat{\mu}_i^{(r)}$ and $\mu_i^{(r)}$ are the estimated and true values of μ_i, respectively, in the rth iteration. The obtained values of RMSEs are averaged over the same groups and the results are reported in Table 8.1.

From Table 8.1, we can observe that the proposed method provides better estimates than three existing methods in almost all cases. As mentioned in the above, ATP method is overfitting in scenario (A) with $\lambda = 0$ while TP method is correctly specified. However, the results show that the performances between ATP and TP are almost the same, which might indicate that the MSE inflation due to overfilling is not serious. The similar observation can be done in scenario (B) with $\lambda = 0$. On the other hand, in the other cases, the proposed ATP method can improve the estimation accuracy of TP method as well as EBP and DE methods, by adaptively estimating the transformation parameter from the data.

3.4.2 Finite sample evaluation of empirical Bayes confidence intervals

We next evaluate the finite sample performances of the empirical Bayes confidence intervals given in Section 3.3. To this end, we considered the following data generating process for population variables Y_{ij}:

$$(2\lambda)^{-1}(Y_{ij}^{\lambda} - Y_{ij}^{-\lambda}) = \beta_0 + \beta_1 X_{ij} + v_i + \varepsilon_{ij}, \quad v_i \sim N(0, \tau^2), \quad \varepsilon_{ij} \sim N(0, \sigma^2),$$

where $j = 1, \ldots, N$ and $i = 1, \ldots, m$ with $N = 200$. We set the true parameter values $\lambda = 0.3$, $\beta_0 = -1$, $\beta_1 = 3$, $\tau = 0.3$, $\sigma = 0.7$, and X_{ij} were initially generated from the uniform distribution on $(1, 2)$, which were fixed through simulation runs. We focused on the same population parameter given in (3.14).

Among the generated $Y_{i1}, \ldots, Y_{iN}$, the first $n = 50$ observations $Y_{i1}, \ldots, Y_{in}$ were used as the sampled data $y_{i1}, \ldots, y_{in}$. Then, based on y_{ij}'s and X_{ij}'s, we computed two types pf confidence intervals for μ_i, naive confidence interval (3.12) and bootstrap calibrated confidence interval (3.13), which are denoted by NCI and BCI, respectively. To evaluate the performances of two confidence intervals, based on $R = 1000$ simulation runs, we computed the empirical coverage probability (CP) and the average length of confidence interval (AL), which are defined as

$$\mathrm{CP}_i = \frac{1}{R}\sum_{r=1}^{R} I(\mu_i^{(r)} \in \mathrm{CI}_i^{(r)}) \quad \text{and} \quad \mathrm{AL}_i = \frac{1}{R}\sum_{r=1}^{R} |\mathrm{CI}_i^{(r)}|,$$

where $\mu_i^{(r)}$ is the true value and $\mathrm{CI}_i^{(r)}$ is NCI or BCI in the rth iteration. In Figure, we show the obtained CP and AL in each area for two cases $m = 20$ and $m = 30$. Concerning CP, the naive method tends to produce shorter confidence intervals, so that the coverage probability is smaller than the nominal level for all areas, which is more serious in case $m = 20$ than $m = 30$. This comes from the accuracy of NCI presented in Theorem 3.2, which mentions that the coverage accuracy of NCI is $O(m^{-1})$. On the other hand, bootstrap method can improve the drawbacks of the naive method, and provides reasonable CP around the nominal level under both $m = 20$ and $m = 30$. The results clearly support the theoretical property given in Theorem 3.3 presenting BCI is second order accurate. Since undervaluation of estimation risk may produce serious problems in practice, we should use the bootstrap method when the number of areas is not large.

Table 3.1: The group-wise averaged values of simulated square root of mean squared errors (RMSE) for four methods, proposed adaptively transformed prediction (ATP) method, Molina and Rao's transformed prediction (TP) method, empirical best prediction (EBP) method without any data transformations, and direct estimator (DE) for eight scenarios. All the values in the table are multiplied by 100.

		Area sample size n_i				
Scenario	Method	20	40	60	80	100
(A) $\lambda = 0$	ATP	3.65	2.68	2.19	1.86	1.62
	TP	3.64	2.68	2.18	1.85	1.62
	EBP	5.09	4.09	3.34	2.83	2.24
	DE	7.77	4.44	5.08	2.47	2.22
(A) $\lambda = 0.2$	ATP	3.58	2.65	2.17	1.83	1.59
	TP	3.74	2.80	2.31	1.95	1.68
	EBP	4.89	4.05	3.29	2.92	2.25
	DE	7.55	4.39	4.91	2.43	2.20
(A) $\lambda = 0.4$	ATP	3.36	2.53	2.07	1.77	1.51
	TP	3.90	3.03	2.52	2.16	1.79
	EBP	3.78	3.02	2.47	2.16	1.74
	DE	6.99	4.26	4.51	2.34	2.09
(B) $\lambda = 0$	ATP	4.58	3.33	2.81	2.31	2.01
	TP	4.58	3.33	2.80	2.31	2.01
	EBP	8.92	7.38	6.10	5.76	4.47
	DE	8.74	6.37	5.79	3.75	2.60
(B) $\lambda = 0.2$	ATP	4.33	3.42	2.85	2.28	1.95
	TP	4.56	3.61	3.03	2.45	2.09
	EBP	6.23	5.08	4.52	3.50	3.01
	DE	8.24	6.45	5.73	3.64	2.58
(B) $\lambda = 0.4$	ATP	4.13	3.25	2.70	2.19	1.93
	TP	4.73	3.88	3.23	2.71	2.30
	EBP	4.68	3.72	3.22	2.61	2.26
	DE	7.82	5.93	5.33	3.46	2.56
(C)	ATP	4.90	3.63	2.96	2.41	2.17
	TP	5.02	3.69	3.03	2.47	2.20
	EBP	6.78	5.74	4.36	3.27	3.11
	DE	8.67	5.31	4.16	4.07	3.05
(D)	ATP	4.54	3.44	2.98	2.53	2.03
	TP	5.05	4.04	3.48	2.97	2.36
	EBP	5.25	4.20	3.38	2.90	2.32
	DE	9.85	5.76	4.74	3.60	3.45

3.4.3 Example: poverty mapping in Spain

We applied the proposed method to estimation of poverty indicators in Spanish provinces, using the synthetic income data available in `sae` package (Molina and Marhuenda, 2015) in `R`

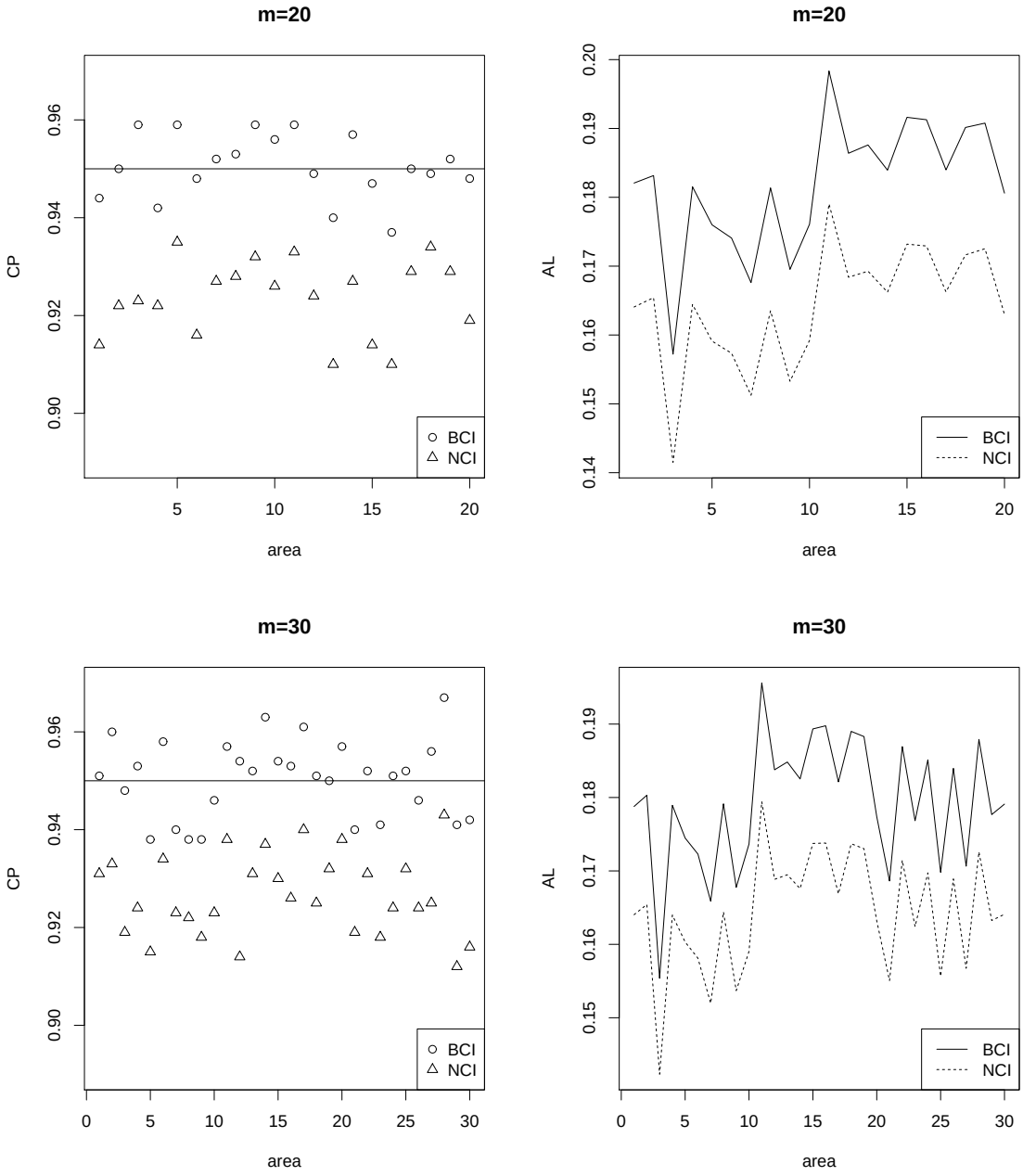

Figure 3.1: Simulated coverage probability (CP) and average length (AL) of two confidence intervals, naive confidence interval (NCI) and bootstrap calibrated confidence interval (BCI) for $m = 20$ (upper) and $m = 30$ (lower).

language, in which the equalized annual net income are given. The similar data set was used in Molina and Rao (2010) and Molina et al. (2014). As auxiliary variables, we considered the indicators of the five quinquennial groupings of the variable age, the indicator of having Spanish nationality, the indicators of the three levels of the variable education level, and the

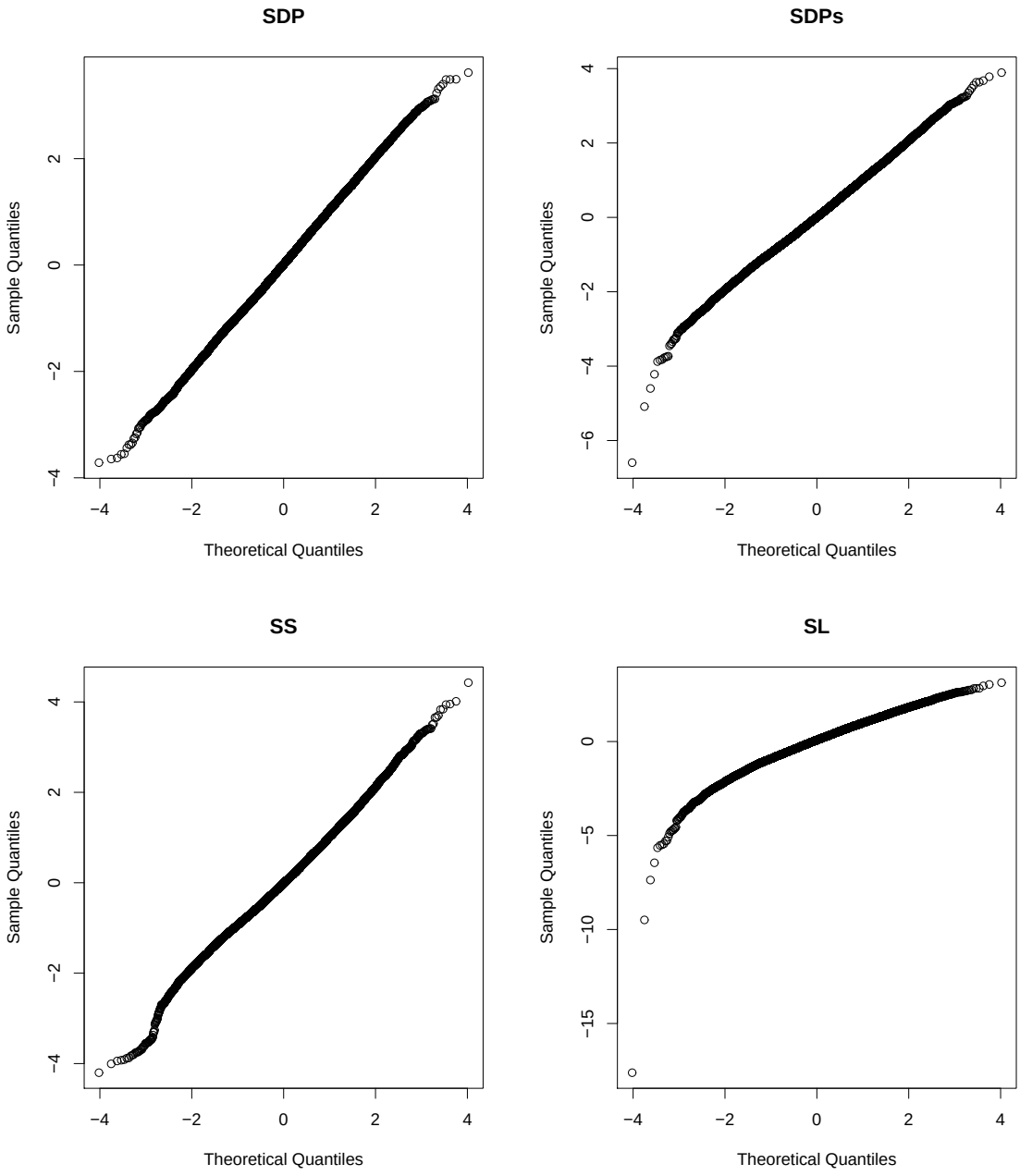

Figure 3.2: QQ-plots of standardized residuals in four models.

indicators of the three categories of the variable employment, with categories unemployed, employed and inactive. For each auxiliary variable, one of the categories was considered as base reference, omitting the corresponding indicator and then including an intercept in the model. The poverty measures we focused on were the FGT poverty measures (Foster et al.,

1984):

$$T(x) = \left(\frac{x-z}{z}\right)^{\alpha} I(x < z),$$

where z is a fixed poverty line, and it corresponds to poverty incidence or head count ratio ($\alpha = 0$), poverty gap ($\alpha = 1$) and poverty severalty ($\alpha = 2$). In this example, we focused on poverty ratio ($\alpha = 0$), and we set z as the 0.6 times the median of incomes. Let E_{ij} be the income of jth individual in ith area. Such data are available for $m = 52$ areas and the sample sizes are are ranging from 20 to 1420. Since the small portion of E_{ij} take negative values, we assume the nested error regression model with shifted-DPT:

$$\text{SDP:} \quad (2\lambda)^{-1}\left\{(E_{ij}+c)^{\lambda} - (E_{ij}+c)^{-\lambda}\right\} = \boldsymbol{x}_{ij}^t\boldsymbol{\beta} + v_i + \varepsilon_{ij}, \tag{3.15}$$

noting that the model has two transformation parameters λ and c. We also considered two submodel of (3.15). In both models, we set $c = c^* \equiv \min(E_{ij}) + 1$ to ensure that $E_{ij} + c^*$ is positive for all (i, j). The first submodel is denied by putting $c = c^*$ in (3.15), which is referred to SDP-s. The second sub-model is the shifted-log transformation model:

$$\text{SL:} \quad \log(E_{ij}+c^*) = \boldsymbol{x}_{ij}^t\boldsymbol{\beta} + v_i + \varepsilon_{ij}, \tag{3.16}$$

which has no longer parameters and was used in Molina and Rao (2010). Finally, we also applied the model with sinh-arcsinh transformation presented in Section 3.2.3:

$$\text{SS:} \quad \sinh(b\sinh^{-1}(E_{ij}) - a) = \boldsymbol{x}_{ij}^t\boldsymbol{\beta} + v_i + \varepsilon_{ij}, \tag{3.17}$$

which has two transformation parameter a and b.

By maximizing the profile likelihood function of transformation parameters, we obtained as follows:

$$\begin{aligned}
&\text{(SDP)} \ \ \widehat{\lambda} = 0.090 \ (1.99 \times 10^{-3}), \quad \widehat{c} = 4319 \ (170.69)\\
&\text{(SDP-s)} \ \ \widehat{\lambda} = 0.290 \ (8.18 \times 10^{-4})\\
&\text{(SS)} \ \widehat{a} = -0.584 \ (8.06 \times 10^{-4}), \quad \widehat{b} = 0.463 \ (1.55 \times 10^{-6}),
\end{aligned}$$

where the values in the parentheses are the corresponding standard errors calculated from the Fisher information matrix given in Theorem 3.1. From the above result, it can be observed that the approximate 95% confidence intervals of the transformation parameter λ in SDP and SDP-s are bounded from 0, which means that the log-transformed model would be inappropriate. Moreover, we computed AIC and BIC based on the maximum marginal likelihood, and the results are given in Table 3.2 in which the values scaled by the number of sampled units ($N = 17199$) are reported. The results show that the SDP fits the best among the four models in terms of both AIC and BIC while the SL model fits the worst. Hence, the use of parametric transformation can improve AIC and BIC in this application. To see the fitting of the models in terms of normality assumption of the error terms, we computed the standardized residuals defined as

$$r_{ij} = \frac{\widehat{H}(y_{ij}) - \boldsymbol{x}_{ij}^t\widehat{\boldsymbol{\beta}}}{\sqrt{\widehat{\tau}^2 + \widehat{\sigma}^2}}, \quad j = 1, \ldots, n_i, \ \ i = 1, \ldots, m,$$

where $\widehat{H}$ is the estimated transformation function, noting that r_{ij}'s asymptotically follow the standard normal distributions if the assumed model is correctly specified. In Figure 3.2, we shows QQ-plots of r_{ij}'s of the four models. We can observe that the normality assumptions in the three models with parametric transformations, SDP, SDP-s and SS, seem plausible from Figure 3.2. However, the QQ-plot for SL shows that the distribution of standardized residuals is skewed and the normality assumption would not be appropriate.

Finally, we calculated the estimated values of the poverty rates μ_i from the direct estimator (DE), and four model based methods. For computing the empirical best predictor of μ_i, we used 100 random samples for Monte Carlo integration. The obtained values are given in Table 3.3 with the empirical Bayes confidence intervals of μ_i. It can be seen that the direct estimator produces quite different estimates of μ_i from the model based methods in Avila and Sevilla. We can also observe that SL method tend to produce larger estimates than the other model based methods. However, from AIC and BIC values and QQ-plot in Figure 3.2, the validity of SL method is highly doubtful in this case, so that the predicted values given in Table 3.3 would not be reliable. As shown in Table 3.3, the use of different transformation function leads to significantly different predicted values of μ_i. Hence, it would be valuable to select an adequate transformation function by estimating transformation parameters based on the sampled data.

Table 3.2: AIC and BIC of four models. The values are scaled by the number of sampled units ($N = 17199$).

	SDP	SDP-s	SS	SL
AIC	20.2241	20.2260	20.2415	20.2883
BIC	20.2305	20.2318	20.2478	20.2937

Table 3.3: Estimated percent poverty rates from the direct estimator (DE) and four model based methods in five provinces. The empirical Bayes confidence intervals are given in the parenthesis.

area	n_i	DE	SDP	SDP-s	SS	SL
Avila	58	8.62	17.81	18.16	18.47	18.58
			(12.33, 23.85)	(12.71, 23.65)	(13.97, 25.28)	(14.36, 24.19)
Tarragona	134	29.17	26.08	26.39	26.59	28.07
			(24.17, 28.34)	(24.30, 28.21)	(24.42 ,28.34)	(25.96, 30.42)
Santander	434	29.31	32.43	33.06	33.03	35.91
			(28.51, 36.00)	(30.19, 35.80)	(30.37, 35.95)	(32.41, 39.02)
Sevilla	482	5.00	25.57	26.26	26.22	27.31
			(23.63, 28.47)	(23.72 ,28.46)	(23.89, 29.29)	(24.88, 29.76)
Oviedo	803	33.33	36.67	37.26	37.80	40.69
			(29.98, 42.72)	(32.55, 42.62)	(31.33, 43.96)	(34.65, 46.66)

3.5 Technical Issues

3.5.1 Proof of Theorem 3.1

From the likelihood function (3.7), its first order derivatives are given by

$$\frac{\partial L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{m} \boldsymbol{X}_i^t \boldsymbol{\Sigma}_i^{-1} \boldsymbol{z}_i, \quad \frac{\partial L}{\partial \tau^2} = -\frac{1}{2}\sum_{i=1}^{m} \mathbf{1}_{n_i}^t \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} - \frac{1}{2}\sum_{i=1}^{m} \boldsymbol{z}_i^t \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^t \boldsymbol{\Sigma}_i^{-1} \boldsymbol{z}_i$$
$$\frac{\partial L}{\partial \sigma^2} = -\frac{1}{2}\sum_{i=1}^{m} \mathrm{tr}(\boldsymbol{\Sigma}_i^{-1}) - \frac{1}{2}\sum_{i=1}^{m} \boldsymbol{z}_i^t \boldsymbol{\Sigma}_i^{-2} \boldsymbol{z}_i,$$
$$\frac{\partial L}{\partial \lambda} = -\sum_{i=1}^{m} \boldsymbol{z}_i^t \boldsymbol{\Sigma}_i^{-1} H_\lambda^{(1)}(y_i) + \sum_{i=1}^{m}\sum_{j=1}^{n_i} \frac{\partial}{\partial \lambda} \log H_\lambda'(y_{ij}),$$

where $\boldsymbol{z}_i = H_\lambda(y_i) - \boldsymbol{X}_i\boldsymbol{\beta}$. Since $\mathrm{E}[\boldsymbol{z}_i] = \mathbf{0}$, it follows that $\mathrm{E}[\partial^2 L/\partial\boldsymbol{\beta}\partial\tau^2] = \mathrm{E}[\partial^2 L/\partial\boldsymbol{\beta}\partial\sigma^2] = \mathbf{0}$. The other elements of the Fisher information can be obtained by a straightforward calculation. Moreover, under Assumptions 3.1 and 3.2, the each element of the Fisher information matrix is finite, so that the asymptotic normality of $\widehat{\boldsymbol{\phi}}$ follows.

3.5.2 Proof of Theorem 3.2

Let $\boldsymbol{\phi}_0$ is the true values of parameters. It suffices to show that $P(\mu_i \le Q_a(y_i, \widehat{\boldsymbol{\phi}})) = a + O(m^{-1})$ for $a \in (0,1)$. We first note that It holds that

$$P(\mu_i \le Q_a(y_i, \widehat{\boldsymbol{\phi}})) = \mathrm{E}[P(\mu_i \le Q_a(y_i, \widehat{\boldsymbol{\phi}})|\boldsymbol{y}_s)] = \mathrm{E}[F(Q_a(y_i, \widehat{\boldsymbol{\phi}}); y_i, \boldsymbol{\phi}_0)],$$

where $F(\cdot; y_i, \boldsymbol{\phi}_0)$ is a distribution function of μ_i given y_i. Let $G(y_i, \widehat{\boldsymbol{\phi}}, \boldsymbol{\phi}_0) = F(Q_a(y_i, \widehat{\boldsymbol{\phi}}); y_i, \boldsymbol{\phi}_0)$, noting that $0 \le G(y_i, \widehat{\boldsymbol{\phi}}, \boldsymbol{\phi}_0) \le 1$ and $G(y_i, \boldsymbol{\phi}_0, \boldsymbol{\phi}_0) = a$. The Taylor expansion of $G(y_i, \widehat{\boldsymbol{\phi}}, \boldsymbol{\phi}_0)$ shows that

$$\begin{aligned} G(y_i, \widehat{\boldsymbol{\phi}}, \boldsymbol{\phi}_0) = G(y_i, \boldsymbol{\phi}_0, \boldsymbol{\phi}_0) &+ \sum_j G_{\phi_j}(y_i, \boldsymbol{\phi}, \boldsymbol{\phi}_0)\big|_{\boldsymbol{\phi}=\boldsymbol{\phi}_0}(\widehat{\phi}_j - \phi_j) \\ &+ \frac{1}{2}\sum_{j,k} G_{\phi_j\phi_k}(y_i, \boldsymbol{\phi}, \boldsymbol{\phi}_0)\big|_{\boldsymbol{\phi}=\boldsymbol{\phi}_0}(\widehat{\phi}_j - \phi_j)(\widehat{\phi}_k - \phi_k) \\ &+ \frac{1}{6}\sum_{j,k,\ell} G_{\phi_j\phi_k\phi_\ell}(y_i, \boldsymbol{\phi}, \boldsymbol{\phi}_0)\big|_{\boldsymbol{\phi}=\boldsymbol{\phi}^*}(\widehat{\phi}_j - \phi_j)(\widehat{\phi}_k - \phi_k)(\widehat{\phi}_\ell - \phi_\ell), \end{aligned}$$

where $\boldsymbol{\phi}^*$ is on the line connecting $\widehat{\boldsymbol{\phi}}$ and $\boldsymbol{\phi}_0$. Then, it follows that

$$P(\mu_i \le Q_a(y_i, \widehat{\boldsymbol{\phi}})) = \mathrm{E}[G(y_i, \widehat{\boldsymbol{\phi}}, \boldsymbol{\phi}_0)] = a + R_1 + \frac{1}{2}R_2 + \frac{1}{6}R_3,$$

where

$$\begin{aligned} R_1 &= \mathrm{E}\left[G_{\boldsymbol{\phi}}(y_i, \boldsymbol{\phi}, \boldsymbol{\phi}_0)\big|_{\boldsymbol{\phi}=\boldsymbol{\phi}_0}(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0)\right] \\ R_2 &= \sum_{j,k} \mathrm{E}\left[G_{\phi_j\phi_k}(y_i, \boldsymbol{\phi}, \boldsymbol{\phi}_0)\big|_{\boldsymbol{\phi}=\boldsymbol{\phi}_0}(\widehat{\phi}_j - \phi_j)(\widehat{\phi}_k - \phi_k)\right] \\ R_3 &= \sum_{j,k,\ell} \mathrm{E}\left[G_{\phi_j\phi_k\phi_\ell}(y_i, \boldsymbol{\phi}, \boldsymbol{\phi}_0)\big|_{\boldsymbol{\phi}=\boldsymbol{\phi}^*}(\widehat{\phi}_j - \phi_j)(\widehat{\phi}_k - \phi_k)(\widehat{\phi}_\ell - \phi_\ell)\right]. \end{aligned}$$

Using the Cauchy-Schwarz inequality, we have

$$\mathrm{E}\left[G_{\phi_j\phi_k}(y_i,\boldsymbol{\phi},\boldsymbol{\phi}_0)\big|_{\boldsymbol{\phi}=\boldsymbol{\phi}_0}(\widehat{\phi}_j-\phi_j)(\widehat{\phi}_k-\phi_k)\right] \\ \leq \left\{\mathrm{E}[(\widehat{\phi}_j-\phi_j)^4]\right\}^{\frac{1}{4}}\left\{\mathrm{E}[(\widehat{\phi}_k-\phi_k)^4]\right\}^{\frac{1}{4}}\sqrt{\mathrm{E}\left[G_{\phi_j\phi_k}(y_i,\boldsymbol{\phi},\boldsymbol{\phi}_0)^2\big|_{\boldsymbol{\phi}=\boldsymbol{\phi}_0}\right]}.$$

From the asymptotic normality of $\widehat{\boldsymbol{\phi}}$ given in Theorem 3.1, it holds that $\mathrm{E}[|\widehat{\phi}_k-\phi_k|^r] = O(m^{-r/2})$. Moreover, since $0 \leq G(y_i,\boldsymbol{\phi},\boldsymbol{\phi}_0) \leq 1$ and $\boldsymbol{\phi}_0$ is an interior point, it holds $|G(y_i,\boldsymbol{\phi}_1,\boldsymbol{\phi}_0) - G(y_i,\boldsymbol{\phi}_2,\boldsymbol{\phi}_0)| \leq 2$ for all $\boldsymbol{\phi}_1,\boldsymbol{\phi}_2 \in N_{\boldsymbol{\phi}_0}$ with $N_{\boldsymbol{\phi}_0} = \{\boldsymbol{\phi}; \|\boldsymbol{\phi}-\boldsymbol{\phi}_0\| \leq \varepsilon\}$, thereby the partial derivatives of $G(y_i,\boldsymbol{\phi},\boldsymbol{\phi}_0)$ at $\boldsymbol{\phi}=\boldsymbol{\phi}_0$ are bounded. Then, we obtain $R_2 = O(m^{-1})$. Using the similar evaluation, we can show that $R_3 = O(m^{-1})$. Regarding R_1, it is noted that

$$\mathrm{E}\left[G_{\boldsymbol{\phi}}(y_i,\boldsymbol{\phi},\boldsymbol{\phi}_0)\big|_{\boldsymbol{\phi}=\boldsymbol{\phi}_0}(\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi}_0)\right] = \mathrm{E}\left[G_{\boldsymbol{\phi}}(y_i,\boldsymbol{\phi},\boldsymbol{\phi}_0)\mathrm{E}[\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi}_0|y_i]\right].$$

From Lohr and Rao (2009), it holds $\mathrm{E}[\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi}_0|y_i] = m^{-1}\boldsymbol{b}_{\boldsymbol{\phi}} - \boldsymbol{I}_{\boldsymbol{\phi}}^{-1}\partial L_i(y_i,\boldsymbol{\phi}_0)/\partial\boldsymbol{\phi} + o_p(m^{-1})$, where $\sum_{i=1}^m L_i(y_i,\boldsymbol{\phi}_0) \equiv L(\boldsymbol{\phi})$ and $b_{\boldsymbol{\phi}} = \lim_{m\to\infty} m\mathrm{E}[\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi}_0]$ is the asymptotic bias of $\widehat{\boldsymbol{\phi}}$. Hence, we have

$$\mathrm{E}\left[G_{\boldsymbol{\phi}}(y_i,\boldsymbol{\phi},\boldsymbol{\phi}_0)\mathrm{E}[\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi}_0|y_i]\right] \\ = \frac{1}{m}\mathrm{E}\left[G_{\boldsymbol{\phi}}(y_i,\boldsymbol{\phi},\boldsymbol{\phi}_0)\right]\boldsymbol{b}_{\boldsymbol{\phi}} - \mathrm{E}\left[G_{\boldsymbol{\phi}}(y_i,\boldsymbol{\phi},\boldsymbol{\phi}_0)\boldsymbol{I}_{\boldsymbol{\phi}}^{-1}\frac{\partial}{\partial\boldsymbol{\phi}}L_i(y_i;\boldsymbol{\phi}_0)\right] + o(m^{-1}),$$

which is $O(m^{-1})$. Therefore, the proof is completed.

3.5.3 Proof of Theorem 3.3

From the proof of Theorem 3.2, we have

$$F_a(\boldsymbol{\phi}_0) \equiv P(\mu_i \leq Q_a(y_i,\widehat{\boldsymbol{\phi}})) = a + \frac{c(a,\boldsymbol{\phi}_0)}{m} + o(m^{-1}),$$

where $c(a,\boldsymbol{\phi})$ is a smooth function of a and $\boldsymbol{\phi}$. Let a^* and $\widehat{a}^*$ be satisfying $F_{a^*}(\boldsymbol{\phi}_0) = a$ and $F_{\widehat{a}^*}(\widehat{\boldsymbol{\phi}}) = a$, respectively. Then, it holds $\widehat{a}^* - a^* = o(1)$ since $\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi} = o(1)$. From the above expansion, we have

$$\widehat{a}^* - a^* = -\frac{1}{m}\left\{c(\widehat{a}^*,\widehat{\boldsymbol{\phi}}) - c(a^*,\boldsymbol{\phi}_0)\right\} + o(m^{-1}),$$

so that $\widehat{a}^* - a^* = o_p(m^{-1})$. Hence, it follows that

$$P(\mu_i \leq Q_{\widehat{a}^*}(y_i,\widehat{\boldsymbol{\phi}})) = P(\mu_i \leq Q_{a^*}(y_i,\widehat{\boldsymbol{\phi}})) + o(m^{-1}) = a + o(m^{-1}),$$

which completes the proof.

3.5.4 Checking assumptions of transformations

We here check the assumption 3 in Assumption 3.1 for the dual power (DP) transformation (3.9) and sinh-arcsinh (SS) transformation (3.10).

(DP transformation) We first note that $H_\lambda^{-1}(x) = O(x^{1/\lambda})$ as $x \to \infty$. By putting $x = -t$ for $t > 0$, we have

$$H_\lambda^{-1}(x) = (\sqrt{1+\lambda^2 t^2} - \lambda t)^{1/\lambda} = \frac{1}{(\sqrt{1+\lambda^2 t^2} + \lambda t)^{1/\lambda}} = O(t^{-1/\lambda})$$

as $t \to \infty$. A straightforward calculation shows that

$$\frac{\partial H_\lambda(x)}{\partial \lambda} = \frac{x^\lambda \log x + x^{-\lambda} \log x}{2\lambda} + \frac{x^\lambda - x^{-\lambda}}{2\lambda^2},$$

thereby, it follows that

$$\left| \frac{\partial H_\lambda}{\partial \lambda}(H_\lambda^{-1}(x)) \right| = O(|x| \log |x|) + O(|x|^{-1} \log |x|) + O(|x|) + O(|x|^{-1}) = O(|x| \log |x|)$$

as $|x| \to \infty$. Moreover, since

$$\frac{\partial^2 H_\lambda(x)}{\partial \lambda^2} = \frac{x^\lambda (\log x)^2 - x^{-\lambda} (\log x)^2}{2\lambda} - \frac{x^\lambda - x^{-\lambda}}{\lambda^3},$$

the similar evaluation leads to $\left|\partial^2 H_\lambda(w)/\partial \lambda^2\right| = O(|x|(\log |x|)^2)$ as $|x| \to \infty$. Regarding $\partial^2 \log H'_\lambda(x)/\partial \lambda^2$, it holds that

$$\left| \frac{\partial^2 \log H'_\lambda(w)}{\partial \lambda^2} \right| = \left| \frac{4(\log w)^2}{w^2 (w^{\lambda-1} + w^{-\lambda-1})^2} \right| = O((\log |x|)^2 |x|^2)$$

as $|x| \to \infty$, so that the DP transformation satisfies the assumption. When the location parameter is used, namely, $H_{\lambda,c}(x) = \{(x+c)^\lambda - (x+c)^{-\lambda}\}/2\lambda$, it is noted that $\partial^k H_{\lambda,c}(x)/\partial c^k = \partial^k H_{\lambda,c}(x)/\partial x^k$, so that the quite similar evaluation shows that the shifted-DP transformation also satisfies the assumption.

(SS transformation) It follows that

$$\frac{\partial H_{a,b}(x)}{\partial a} = -\cosh(b \sinh^{-1}(x) - a), \qquad \frac{\partial H_{a,b}(x)}{\partial b} = \cosh(b \sinh^{-1}(x) - a) \sinh^{-1}(x).$$

Note that $\sinh^{-1}(x) = O(\log |x|)$ as $|x| \to \infty$, so that $H_{a,b}^{-1}(x) = O(\exp(b^{-1} \log |x|)) = O(|x|^{1/b})$. Then, we have

$$\begin{aligned}
&\frac{\partial H_{a,b}}{\partial a}(H_{a,b}^{-1}(x)) = O(\exp(b \log |x|^{1/b})) = O(|x|), \\
&\frac{\partial H_{a,b}}{\partial b}(H_{a,b}^{-1}(x)) = O(\exp(b \log |x|^{1/b}) \log |x|^{1/b}) = O(|x| \log |x|),
\end{aligned}$$

as $|x| \to \infty$. Moreover, it holds that

$$\begin{aligned}
&\frac{\partial^2 H_{a,b}(x)}{\partial^2 a} = \sinh(b \sinh^{-1}(x) - a), \qquad \frac{\partial^2 H_{a,b}(x)}{\partial^2 b} = \sinh(b \sinh^{-1}(x) - a)\{\sinh^{-1}(x)\}^2 \\
&\frac{\partial^2 H_{a,b}(x)}{\partial a \partial b} = -\sinh(b \sinh^{-1}(x) - a) \sinh^{-1}(x),
\end{aligned}$$

thereby the similar evaluation shows that $\partial^2 H_{a,b}(x)/\partial^2 a = O(|x|)$, $\partial^2 H_{a,b}(x)/\partial^2 b = O(|x|(\log|x|)^2)$ and $\partial^2 H_{a,b}(x)/\partial a \partial b = O(|x|\log|x|)$ as $|x| \to \infty$. On the other hand, a straightforward calculation shows that

$$\frac{\partial}{\partial a}\log H'_{a,b}(x) = \frac{H_{a,b}(x)}{1+H_{a,b}(x)^2}\frac{\partial H_{a,b}(x)}{\partial a}, \qquad \frac{\partial}{\partial b}\log H'_{a,b}(x) = \frac{1}{b} + \frac{H_{a,b}(x)}{1+H_{a,b}(x)^2}\frac{\partial H_{a,b}(x)}{\partial b},$$

which are bounded by the function $\partial H_{a,b}(x)/\partial a$ and $\partial H_{a,b}(x)/\partial b$, respectively. A straightforward calculations show that the second partial derivatives of $\log H'_{a,b}(x)$ are bounded by polynomial functions of the second partial derivatives of $H_{a,b}(x)$ and $H_{a,b}(x)$, thereby the assumption is satisfied.

Chapter 4

Conditional Mean Squared Errors in Mixed Models

4.1 Introduction

Traditionally, the (unconditional) mean squared errors (MSE) have been used for assessing the variability of model based estimators. However, it is criticized that the unconditional MSE do not give us appropriate estimation errors, since it is an integrated measure. Booth and Hobert (1998) suggested the conditional MSE (CMSE) given the data of the small area of interest, and Datta et al. (2011a) and Torabi and Rao (2013) derived second-order unbiased estimators of the conditional MSE in the Fay-Herriot model and nested error regression model, respectively. As pointed out in both papers, the difference between the conditional and unconditional MSEs is small in the model based on normal distribution since it appears in the second-order terms. On the other hand, in the generalized linear mixed models, Booth and Hobert (1998) showed that the difference is significant for distributions far from normality in the sense that the difference is appeared in the first-order. Although the generalized linear mixed models are useful for analyzing count data in small area estimation, it is computationally hard to derive the small area estimator and to evaluate their conditional MSEs, because the marginal likelihood and the small area estimator in the generalized linear mixed model cannot be expressed in closed forms. In fact, we need relatively high dimensional numerical integration to evaluate the conditional MSEs. Another point is the assumption that sample sizes of small areas are large, under which the Laplace approximation can be used to get asymptotically unbiased estimators of the conditional MSEs. However, this assumption is against the situation in small area estimation with small samples sizes.

An alternative model is the mixed model based on the natural exponential families with quadratic variance functions (NEF-QVF) developed in Ghosh and Maiti (2004). The NEF-QVF models include the Fay-Herriot model as well as Poisson-gamma and binomial-beta models, which are extensively used in a variety of applications. The practical advantage compared with generalized linear mixed models is that the Bayes estimator of the small area parameter is the weighed average of a sample mean and a prior mean, so that the estimator is easy to compute without any numerical techniques. However, there has been no literatures concerned with the CMSEs in the NEF-QVF model in spite of their importance. Hence, in this paper, we investigate the CMSE of the mixed models based on the NEF-QVF, and derive

the second-order unbiased estimator of CMSE for practical use.

In Section 4.2, we first provide general strategies of deriving a second order unbiased estimator of CMSE based on the parametric bootstrap and an analytical method based on Taylor expansion. Then, in Section 4.3, we investigate the properties of CMSEs in NEF-QVF mixed models and derive a second order unbiased estimator of CMSEs using the results in 4.2. In Section 4.4, we studies the numerical properties of the CMSE estimator through simulation and empirical studies. All the technical proofs are given in Section 4.5

4.2 Conditional MSE in General Mixed Models

Let y_i be a direct estimate of small area parameter θ_i for $i = 1, \ldots, m$. In this paper, we treat both continuous and discrete cases for y_i and θ_i. We consider the following two stage genera mixed model:

$$y_i|(\theta_i, \boldsymbol{\phi}) \sim f(y_i|\theta_i, \boldsymbol{\phi}), \quad \theta_i|\boldsymbol{\phi} \sim \pi(\theta_i|\boldsymbol{\phi}) \qquad i = 1, \ldots, m,$$

where $\boldsymbol{\phi}$ is a q-dimensional vector of unknown parameters. In the continuous case, the marginal density function of y_i for given $\boldsymbol{\phi}$ and the conditional (or posterior) density function of θ_i given y_i are given by

$$\begin{aligned} m_\pi(y_i|\boldsymbol{\phi}) =& \int f(y_i|\theta_i, \boldsymbol{\phi})\pi(\theta_i|\boldsymbol{\phi})\mathrm{d}\theta_i \\ \pi(\theta_i|y_i, \boldsymbol{\phi}) =& f(y_i|\theta_i, \boldsymbol{\phi})\pi(\theta_i|\boldsymbol{\phi})/m_\pi(y_i|\boldsymbol{\phi}) \end{aligned}$$

and we use the same notations in the discrete case. Then, for $i = 1, \ldots, m$, we consider the problem of estimating (predicting) a scalar quantity $\mu_i(\theta_i, \boldsymbol{\phi})$ of each small area.

For generic estimator $\widehat{\mu}_i$, the risk of the estimator is evaluated the unconditional and conditional MSEs, described as

$$\begin{aligned} \mathrm{MSE}_i =& \mathrm{E}\Big[\big\{\widehat{\mu}_i - \mu_i(\theta_i, \boldsymbol{\phi})\big\}^2\Big], \\ \mathrm{CMSE}_i =& \mathrm{E}\Big[\big\{\widehat{\mu}_i - \mu_i(\theta_i, \boldsymbol{\phi})\big\}^2|y_i\Big]. \end{aligned}$$

Since $y_1, \ldots, y_m$ are independent, the best predictor of $\mu_i(\theta_i, \boldsymbol{\phi})$ in terms of the two kinds of MSEs are the conditional expectation given by

$$\widetilde{\mu}_i \equiv \widetilde{\mu}_i(y_i, \boldsymbol{\phi}) = \mathrm{E}\left[\mu_i(\theta_i, \boldsymbol{\phi})|y_i\right],$$

which corresponds to the Bayes estimator of μ_i. However, the hyperparameter $\boldsymbol{\phi}$ is unknown and $\widetilde{\mu}_i$ is infeasible, we need to estimate $\boldsymbol{\phi}$ from observations $y_1, \ldots, y_m$. Substituting an estimator $\widehat{\boldsymbol{\phi}}$ into $\widetilde{\mu}_i(y_i, \boldsymbol{\phi})$, we obtain an empirical Bayes (EB) estimator $\widehat{\mu}_i \equiv \widetilde{\mu}_i(y_i, \widehat{\boldsymbol{\phi}})$.

For risk evaluation of an empirical Bayes estimator, we here focus on asymptotic evaluations of the CMSE. To this end, we assume the following conditions on the estimator $\widehat{\boldsymbol{\phi}}$ and the Bayes estimator $\widetilde{\mu}_i(y_i, \boldsymbol{\phi})$ for large m:

Assumption 4.1.

(i) The dimension q of $\boldsymbol{\phi}$ is bounded and the estimator $\widehat{\boldsymbol{\phi}}$ satisfies that $(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi})|y_i = O_p(m^{-1/2})$, $\mathrm{E}[\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}|y_i] = O_p(m^{-1})$ *and* $\mathrm{Var}(\widehat{\boldsymbol{\phi}}|y_i) = O_p(m^{-1})$ *for* $i = 1, \ldots, m$.

(ii) For $i = 1, \ldots, m$, $\mu_i(\theta_i, \boldsymbol{\phi}) = O_p(1)$, $\widehat{\mu}_i(y_i, \boldsymbol{\phi}) = O_p(1)$, *and the variances* $\mathrm{Var}(\mu_i(\theta_i, \boldsymbol{\phi})|y_i)$ *and* $\mathrm{Var}(\widetilde{\mu}_i(y_i, \boldsymbol{\phi}))$ *exist.*

(iii) The estimator $\widetilde{\mu}_i(y_i, \boldsymbol{\phi})$ *is continuously differentiable with respect to* $\phi_j, j = 1, \ldots, q$, *and satisfies*

$$\frac{\partial \widetilde{\mu}_i(y_i, \boldsymbol{\phi})}{\partial \phi_j} = O_p(1), \ E\left[\left|\frac{\partial \widetilde{\mu}_i(y_i, \boldsymbol{\phi})}{\partial \phi_j}\right| \Big| y_i\right] < \infty.$$

Under Assumption 4.1, we get a second-order approximation of CMSE of $\widehat{\mu}_i$. Let

$$\begin{aligned} T_{1i}(y_i, \boldsymbol{\phi}) =& \mathrm{Var}(\mu_i(\theta_i, \boldsymbol{\phi})|y_i), \\ T_{2i}(y_i, \boldsymbol{\phi}) =& \mathrm{E}\Big[\Big\{(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi})^t \frac{\partial \widetilde{\mu}_i(y_i, \boldsymbol{\phi})}{\partial \boldsymbol{\phi}}\Big\}^2 \Big| y_i\Big], \end{aligned}$$

where $T_{1i}(y_i, \boldsymbol{\phi})$ is the conditional or posterior variance of $\mu_i(\theta_i, \boldsymbol{\phi})$. It is noted that $T_{1i}(y_i, \boldsymbol{\phi}) = O_p(1)$ and $T_{2i}(y_i, \boldsymbol{\phi}) = O_p(m^{-1})$ under Assumption 4.1.

Theorem 4.1. *Under Assumption 4.1, the CMSE of the empirical Bayes estimator* $\widehat{\mu}_i$ *is approximated as*

$$\mathrm{CMSE}_i = T_{1i}(y_i, \boldsymbol{\phi}) + T_{2i}(y_i, \boldsymbol{\phi}) + o_p(m^{-1}).$$

Proof. Since $\mathrm{E}[\mu_i - \widetilde{\mu}_i|y_i] = 0$, it is observed that

$$\mathrm{CMSE}_i = \mathrm{E}[(\mu_i - \widetilde{\mu}_i + \widetilde{\mu}_i - \widehat{\mu}_i)^2|y_i] = \mathrm{E}[(\mu_i - \widetilde{\mu}_i)^2|y_i] + \mathrm{E}[(\widehat{\mu}_i - \widetilde{\mu}_i)^2|y_i], \tag{4.1}$$

and that $\mathrm{E}[(\mu_i - \widetilde{\mu}_i)^2|y_i] = \mathrm{Var}(\mu_i|y_i) = T_{1i}(y_i, \boldsymbol{\phi})$. It is noted that

$$\widehat{\mu}_i = \widetilde{\mu}_i + \Big\{\frac{\partial \widetilde{\mu}_i(y_i, \boldsymbol{\phi}^*)}{\partial \boldsymbol{\phi}}\Big\}^t (\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}),$$

where $\boldsymbol{\phi}^*$ is between $\boldsymbol{\phi}$ and $\widehat{\boldsymbol{\phi}}$. Since $(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}) \mid y_i = O_p(m^{-1/2})$, we obtain

$$\mathrm{E}[(\widehat{\mu}_i - \widetilde{\mu}_i)^2|y_i] = \mathrm{E}\Big[\Big\{(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi})^t \frac{\partial \widehat{\mu}_i(y_i, \boldsymbol{\phi})}{\partial \boldsymbol{\phi}}\Big\}^2 \Big| y_i\Big] + o_p(m^{-1}),$$

which shows Theorem 4.1. □

We next derive second order unbiased estimators of T_1 and T_2, which result in a second order unbiased estimator of CMSE. As seen from Theorem 4.1, the order of $T_{2i}(y_i, \boldsymbol{\phi})$ is $O_p(m^{-1})$, so that we can estimate $T_{2i}(y_i, \boldsymbol{\phi})$ by the plug-in estimator $T_{2i}(y_i, \widehat{\boldsymbol{\phi}})$ unbiasedly up to second order. For estimation of $T_{1i}(y_i, \boldsymbol{\phi})$, the naive estimator $T_{1i}(y_i, \widehat{\boldsymbol{\phi}})$ has a second order bias because $T_{1i}(y_i, \boldsymbol{\phi}) = O_p(1)$. It is observed that

$$\mathrm{E}[T_{1i}(y_i, \widehat{\boldsymbol{\phi}})|y_i] = T_{1i}(y_i, \boldsymbol{\phi}) + T_{11i}(y_i, \boldsymbol{\phi}) + T_{12i}(y_i, \boldsymbol{\phi}) + o_p(m^{-1}), \tag{4.2}$$

where

$$T_{11i}(y_i, \boldsymbol{\phi}) = \Big\{\frac{\partial T_{1i}(y_i, \boldsymbol{\phi})}{\partial \boldsymbol{\phi}}\Big\}^t \mathrm{E}[(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi})|y_i]$$

and

$$T_{12i}(y_i,\boldsymbol{\phi}) = \frac{1}{2}\mathrm{tr}\Big[\Big\{\frac{\partial^2 T_{1i}(y_i,\boldsymbol{\phi})}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}^t}\Big\}\mathrm{E}\left[(\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi})(\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi})^t|y_i\right]\Big].$$

It is noted that $T_{11i}(y_i,\boldsymbol{\phi}) = O_p(m^{-1})$ and $T_{12i}(y_i,\boldsymbol{\phi}) = O_p(m^{-1})$ under Assumption 4.1.

[Analytical method] It follows from (4.2) that a second order unbiased estimator of CMSE is given by

$$\widehat{\mathrm{CMSE}}_i = T_{1i}(y_i,\widehat{\boldsymbol{\phi}}) - T_{11i}(y_i,\widehat{\boldsymbol{\phi}}) - T_{12i}(y_i,\widehat{\boldsymbol{\phi}}) + T_{2i}(y_i,\widehat{\boldsymbol{\phi}}). \tag{4.3}$$

Theorem 4.2. *Under Assumption 4.1, the estimator* (4.3) *is a second-order unbiased estimator of CMSE in the sense that*

$$\mathrm{E}\big[\widehat{\mathrm{CMSE}}_i|y_i\big] = \mathrm{CMSE}_i + o_p(m^{-1}).$$

As explained in Section 4.3, in the mixed model based on NEF-QVF, we can provide analytical expressions for T_{11i} and T_{12i}, whereby we obtain a second order unbiased estimator in a closed form. In general, however, it is hard to get analytical expressions for T_{11i} and T_{12i}. In this case, as given below, the parametric bootstrap method helps us provide a feasible second order unbiased estimator of CMSE.

[Parametric bootstrap method] Since y_i is fixed, a bootstrap sample is generated from

$$y_j^*|(\theta_j^*,\widehat{\boldsymbol{\phi}}) \sim f(y_j^*|\theta_j^*,\widehat{\boldsymbol{\phi}}) \qquad j\neq i,\ j=1,\ldots,m,$$

where θ_j^*'s are mutually independently distributed as $\theta_j^*|\widehat{\boldsymbol{\phi}} \sim \pi(\theta_j^*|\widehat{\boldsymbol{\phi}})$. Noting that y_i is fixed, we construct the estimator $\widehat{\boldsymbol{\phi}}_{(i)}^*$ from the bootstrap sample

$$y_1^*,\ldots,y_{i-1}^*,y_i,y_{i+1}^*,\ldots,y_m^*$$

with the same technique as used to obtain the estimator $\widehat{\boldsymbol{\phi}}$. Let $\mathrm{E}_*[\cdot|y_i]$ be the expectation with regard to the bootstrap sample. A second order unbiased estimator of $T_{1i}(y_i,\boldsymbol{\phi})$ is given by

$$\overline{T}_{1i}(y_i,\widehat{\boldsymbol{\phi}}) = 2T_{1i}(y_i,\widehat{\boldsymbol{\phi}}) - \mathrm{E}_*\left[T_{1i}(y_i,\widehat{\boldsymbol{\phi}}_{(i)}^*)|y_i\right].$$

Then, it can be verified that $\mathrm{E}[\overline{T}_{1i}(y_i,\widehat{\boldsymbol{\phi}})|y_i] = T_{1i}(y_i,\boldsymbol{\phi}) + o_p(m^{-1})$. In fact, from (4.2), it is noted that

$$\mathrm{E}[T_{1i}(y_i,\widehat{\boldsymbol{\phi}})|y_i] = T_{1i}(y_i,\boldsymbol{\phi}) + d_i(y_i,\boldsymbol{\phi}) + o_p(m^{-1}),$$

where $d_i(y_i,\boldsymbol{\phi}) = T_{11i}(y_i,\boldsymbol{\phi}) + T_{12i}(y_i,\boldsymbol{\phi})$. This implies that $\mathrm{E}_*\left[T_{1i}(y_i,\widehat{\boldsymbol{\phi}}_{(i)}^*)|y_i\right] = T_{1i}(y_i,\widehat{\boldsymbol{\phi}}) + d_i(y_i,\widehat{\boldsymbol{\phi}}) + o_p(m^{-1})$. Since $d_i(y_i,\boldsymbol{\phi})$ is continuous in $\boldsymbol{\phi}$ and $d_i(y_i,\boldsymbol{\phi}) = O_p(m^{-1})$, one gets $\mathrm{E}[\overline{T}_{1i}(y_i,\widehat{\boldsymbol{\phi}})|y_i] = T_{1i}(y_i,\boldsymbol{\phi}) + o_p(m^{-1})$.

For $T_{2i}(y_i,\boldsymbol{\phi})$, from (4.1), it is estimated via parametric bootstrap method as

$$T_{2i}^*(y_i,\widehat{\boldsymbol{\phi}}) = \mathrm{E}^*\big[\{\widetilde{\mu}_i(y_i,\widehat{\boldsymbol{\phi}}_{(i)}^*) - \widetilde{\mu}_i(y_i,\widehat{\boldsymbol{\phi}})\}^2\big|y_i\big].$$

It is noted that the estimator $T_{2i}^*(y_i,\widehat{\boldsymbol{\phi}})$ is always available although an analytical expression of $T_{2i}(y_i,\boldsymbol{\phi})$ is not necessarily available. Combining the above results yields the estimator

$$\widehat{\mathrm{CMSE}}_i^* = \overline{T}_{1i}(y_i,\widehat{\boldsymbol{\phi}}) + T_{2i}^*(y_i,\widehat{\boldsymbol{\phi}}). \tag{4.4}$$

Theorem 4.3. *Under Assumption 4.1, the estimator* (4.4) *is a second-order unbiased estimator of CMSE in the sense that*

$$\mathrm{E}[\widehat{\mathrm{CMSE}}_i^*|y_i] = \mathrm{CMSE}_i + o_p(m^{-1}).$$

4.3 Applications to NEF-QVF

We now consider the mixed models based on natural exponential families with quadratic variance functions (NEF-QVF). The NEF-QVF mixed models in context of small area estimation were proposed by Ghosh and Maiti (2004), in which a second order unbiased estimator of the unconditional MSE was used for qualify the risk of an empirical Bayes estimator. As mentioned before, the CMSE is more preferable than the MSE as a risk measure in the context of small area estimation, we here apply the results in the previous section to provide a second-order approximation and its unbiased estimator for the CMSE.

4.3.1 Empirical Bayes estimator in NEF-QVF

Let $y_1, \ldots, y_m$ be mutually independent random variables where the conditional distribution of y_i given θ_i and the marginal distribution of θ_i belong to the the following natural exponential families:

$$\begin{aligned} y_i|\theta_i \sim & f(y_i|\theta_i) = \exp[n_i(\theta_i y_i - \psi(\theta_i)) + c(y_i, n_i)], \\ \theta_i|\nu, m_i \sim & \pi(\theta_i|\nu, m_i) = \exp[\nu(m_i\theta_i - \psi(\theta_i))]C(\nu, m_i), \end{aligned} \tag{4.5}$$

where n_i is a known scalar parameter and ν is an unknown scalar hyperparameter. Let $\boldsymbol{y} = (y_1, \ldots, y_m)^t$ and $\boldsymbol{\phi} = (\theta_1, \ldots, \theta_m)^t$. The function $f(y_i|\theta_i)$ is the regular one-parameter exponential family and the function $\pi(\theta_i|\nu, m_i)$ is the conjugate prior distribution. Define μ_i by

$$\mu_i = \mathrm{E}[y_i|\theta_i] = \psi'(\theta_i),$$

which is the conditional expectation of y_i given θ_i, where $\psi'(x) = d\psi(x)/dx$. Assume that $\psi''(\theta_i) = Q(\mu_i)$ for $\psi''(x) = d^2\psi(x)/dx^2$, namely,

$$\mathrm{Var}(y_i|\theta_i) = \frac{\psi''(\theta_i)}{n_i} = \frac{Q(\mu_i)}{n_i},$$

where $Q(x) = v_0 + v_1 x + v_2 x^2$ for known constants v_0, v_1 and v_2 which are not simultaneously zero. This means that given θ_i, the conditional variance $\mathrm{Var}(y_i|\theta_i)$ is a quadratic function of the conditional expectation $\mathrm{E}[y_i|\theta_i]$. This is the natural exponential family with the quadratic variance function (NEF-QVF) studied by Morris (1982, 1983). Similarly, the mean and variance of the prior distribution are given by

$$\mathrm{E}[\mu_i|m_i, \nu] = m_i, \quad \mathrm{Var}(\mu_i|m_i, \nu) = \frac{Q_i(m_i)}{\nu - v_2}.$$

In our settings, we consider the link given by

$$m_i = \psi'(\boldsymbol{x}_i^t\boldsymbol{\beta}), \quad i = 1, \ldots, m,$$

where $\boldsymbol{x}_i$ is a $p \times 1$ vector of explanatory variables and $\boldsymbol{\beta}$ is a $p \times 1$ unknown common vector of regression coefficients. Then, the unknown parameters $\boldsymbol{\phi}$ in the previous section correspond to $\boldsymbol{\phi}^t = (\boldsymbol{\beta}^t, \nu)$. The joint probability density (or mass) function of (y_i, θ_i) can be expressed as

$$f(y_i|\theta_i)\pi(\theta_i|\nu, m_i) = \pi(\theta_i|y_i, \nu) f_\pi(y_i|\nu, m_i),$$

where $\pi(\theta_i|y_i, \nu)$ is the conditional (or posterior) density function of θ_i given y_i, and $f_\pi(y_i|\nu, m_i)$ is the marginal density function of y_i. These density (or mass) functions are written as

$$\begin{aligned} \pi(\theta_i|y_i, \nu, m_i) =& \exp[(n_i + \nu)(\widetilde{\mu}_i\theta_i - \psi(\theta_i))]C(n_i + \nu, \widetilde{\mu}_i), \\ f_\pi(y_i|\nu, m_i) =& \frac{C(\nu, m_i)}{C(n_i + \nu, \widetilde{\mu}_i)} \exp[c(y_i, n_i)], \end{aligned} \tag{4.6}$$

where $\widetilde{\mu}_i$ is the posterior expectation of μ_i, namely, $\widetilde{\mu}_i = \mathrm{E}[\mu_i|y_i; \boldsymbol{\phi}]$, given by

$$\widetilde{\mu}_i \equiv \widetilde{\mu}_i(y_i, \boldsymbol{\phi}) = \frac{n_i y_i + \nu m_i}{n_i + \nu}, \tag{4.7}$$

which corresponds to the Bayes estimator of μ_i in the Bayesian context when ν and m_i are known. As shown in Ghosh and Maiti (2008),

$$\mathrm{E}[y_i] = m_i, \qquad \mathrm{Var}(y_i) = Q_i(m_i)\phi_i, \qquad \mathrm{Cov}(y_i, \mu_i) = Q_i(m_i)/(\nu - v_2),$$

for $\phi_i = (1 + \nu/n_i)/(\nu - v_2)$. Using these observations, Ghosh and Maiti (2008) showed that the Bayes estimator $\widetilde{\mu}_i$ given in (4.7) is the best linear unbiased predictor of μ_i under the squared loss.

Concerning the estimation of unknown hyperparameter $\boldsymbol{\phi}$, Ghosh and Maiti (2004) suggested the use of the optimal estimating equations developed in Godambe and Thompson (1989). Let $\boldsymbol{g}_i = (g_{1i}, g_{2i})^t$ for $g_{1i} = y_i - m_i$ and $g_{2i} = (y_i - m_i)^2 - \phi_i Q_i(m_i)$. Moreover, let

$$\boldsymbol{D}_i^t = Q_i(m_i) \begin{pmatrix} \boldsymbol{x}_i & Q_i'(m_i)\phi_i\boldsymbol{x}_i \\ 0 & -(1 + v_2/n_i)(\nu - v_2)^{-2} \end{pmatrix}, \quad \boldsymbol{\Sigma}_i \equiv \mathrm{Cov}(\boldsymbol{g}_i) = \begin{pmatrix} \mu_{2i} & \mu_{3i} \\ \mu_{3i} & \mu_{4i} - \mu_{2i}^2 \end{pmatrix},$$

and $|\boldsymbol{\Sigma}_i| = \mu_{4i}\mu_{2i} - \mu_{2i}^3 - \mu_{3i}^2$, where $\mu_{ri} = \mathrm{E}[(y_i - m_i)^r]$, $r = 1, 2, \ldots$, and exact expressions of μ_{2i}, μ_{3i} and μ_{4i} are given below. Then, Ghosh and Maiti (2008) derived the estimating equations of the form $\sum_{i=1}^m \boldsymbol{D}_i^t \boldsymbol{\Sigma}_i^{-1} \boldsymbol{g}_i = \boldsymbol{0}$, which are written as

$$\begin{aligned} &\sum_{i=1}^m \frac{1}{|\boldsymbol{\Sigma}_i|}\Big[\{\mu_{4i} - \mu_{2i}^2 - \mu_{3i}\phi_i Q_i'(m_i)\}g_{1i} + \{\mu_{2i}\phi_i Q_i'(m_i) - \mu_{3i}\}g_{2i}\Big] Q_i(m_i)\boldsymbol{x}_i = \boldsymbol{0}, \\ &\sum_{i=1}^m \frac{1}{|\boldsymbol{\Sigma}_i|}\{\mu_{2i}g_{2i} - \mu_{3i}g_{1i}\}Q_i(m_i)(1 + v_2/n_i)(\nu - v_2)^{-2} = 0. \end{aligned} \tag{4.8}$$

The resulting estimator of $\boldsymbol{\phi}$ is here called the GT-estimator and denoted by $\widehat{\boldsymbol{\phi}}_{\mathrm{GT}}$. The equations can be solved numerically. In our numerical investigation, we used the `optim` function in 'R' to solve the estimating equations by minimizing the sums of squares of the estimating functions. The exact moments $\mu_{ri} = \mathrm{E}[(y_i - m_i)^r]$, $r = 1, 2, 3, 4$, are obtain from Theorem 1 of Ghosh and Maiti (2004) as

$$\mu_{2i} = \frac{Q(m_i)(\nu/n_i + 1)}{\nu - v_2}, \quad \mu_{3i} = \frac{Q(m_i)Q'(m_i)(\nu/n_i + 1)(\nu/n_i + 2)}{(\nu - v_2)(\nu - 2v_2)},$$

and

$$\begin{aligned}\mu_{4i} =&(d_i+1)(2d_i+1)(3d_i+1)\mathrm{E}[(\mu_i-m_i)^4]+\frac{6}{n_i}Q_i'(m_i)(d_i+1)(2d_i+1)\mathrm{E}[(\mu_i-m_i)^3]\\&+\frac{d_i+1}{n_i^2}\big[7\{Q'(m_i)\}^2+2n_i(4d_i+3)Q(m_i)\big]\mathrm{E}[(\mu_i-m_i)^2]\\&+\frac{1}{n_i^3}Q(m_i)\big[n_i(2d_i+3)Q(m_i)+\{Q'(m_i)\}^2\big],\end{aligned}$$

for $d_i = v_2/n_i$. The expressions of the moments of μ_i are obtained given in Kubokawa et al. (2014) as $\mathrm{E}[(\mu_i-m_i)^2] = Q(m_i)/(\nu-v_2)$, $\mathrm{E}[(\mu_i-m_i)^3] = 2Q(m_i)Q'(m_i)/(\nu-v_2)(\nu-2v_2)$ and

$$\mathrm{E}\left[(\mu_i-m_i)^4\right] = \frac{3Q(m_i)\left[(\nu-v_2)Q(m_i)+2\left\{Q'(m_i)\right\}^2\right]}{(\nu-v_2)(\nu-2v_2)(\nu-3v_2)}.$$

Using these expressions, the estimating equation (4.8) is completed.

An alternative method for estimating $\boldsymbol{\phi}$ is the maximum likelihood (ML) estimator. Since a closed expression of the marginal distribution of $\boldsymbol{y}$ is given in (4.6) in the NEF-QVF mixed model, the ML estimator of $\boldsymbol{\phi}$ is defined as

$$\widehat{\boldsymbol{\phi}}_{\mathrm{ML}} = \mathrm{argmax}_{\boldsymbol{\phi}}\left\{\sum_{i=1}^m \log\frac{C(\nu,m_i)}{C(n_i+\nu,\widetilde{\mu}_i(y_i,\boldsymbol{\phi}))}\right\}.$$

When the parameter $\boldsymbol{\phi}$ is estimated by the GT-estimator $\widehat{\boldsymbol{\phi}} = \widehat{\boldsymbol{\phi}}_{\mathrm{GT}}$ or the ML-estimator $\widehat{\boldsymbol{\phi}} = \widehat{\boldsymbol{\phi}}_{\mathrm{ML}}$, we can construct the estimator $\widehat{m}_i = \psi'(\boldsymbol{x}_i^t\widehat{\boldsymbol{\beta}})$ of m_i. Substituting $\widehat{m}_i$ and $\widehat{\nu}$ into (4.7), we finally get the empirical Bayes estimator of μ_i:

$$\widehat{\mu}_i \equiv \widetilde{\mu}_i(y_i,\widehat{\boldsymbol{\phi}}) = \frac{n_iy_i+\widehat{\nu}\widehat{m}_i}{n_i+\widehat{\nu}}. \tag{4.9}$$

4.3.2 Evaluation of the CMSE

Our interest is evaluating the CMSE of $\widehat{\mu}_i$ given in (4.9). Since the second-order approximation of the CMSE is given in Theorem 4.1, we need to evaluate the first and second order terms $T_{1i}(y_i,\boldsymbol{\phi})$ and $T_{2i}(y_i,\boldsymbol{\phi})$ in the CMSE. For the first order term, it is easy to see that

$$T_{1i}(y_i,\boldsymbol{\phi}) = \mathrm{Var}(\mu_i(\theta_i,\boldsymbol{\phi})|y_i) = \frac{Q(\widehat{\mu}_i(y_i,\boldsymbol{\phi}))}{n_i+\nu-v_2}, \qquad i=1,\ldots,m, \tag{4.10}$$

which is $O_p(1)$. For the second order term, unfortunately, we do not have an analytical expression of $T_{2i}(y_i,\boldsymbol{\phi})$ when we use the ML-estimator $\widehat{\boldsymbol{\phi}}_{\mathrm{ML}}$ for $\widehat{\boldsymbol{\phi}}$. But, the parametric bootstrap method given in Theorem 4.3 enables us to obtain a second order unbiased estimator of the CMSE. When the GT-estimator $\widehat{\boldsymbol{\phi}}_{\mathrm{GT}}$ is used for $\boldsymbol{\phi}$, on the other hand, we can derive an analytical expression of $T_{2i}(y_i,\boldsymbol{\phi})$, which yields closed forms of the second-order approximation of the CMSE and the asymptotically unbiased estimator of the CMSE. Thus, in the rest of this subsection, we focus on derivation of analytical expressions for the CMSE when the GT-estimator $\widehat{\boldsymbol{\phi}}_{\mathrm{GT}}$ is used for $\boldsymbol{\phi}$.

We begin by giving a stochastic expansion and conditional moments of $\widehat{\boldsymbol{\phi}}_{\rm GT}$ which is the solution of the estimating equations (4.8). We use the following notations:

$$\boldsymbol{s}_m = \sum_{i=1}^{m} \boldsymbol{D}_i^t \boldsymbol{\Sigma}_i^{-1} \boldsymbol{g}_i, \qquad \boldsymbol{U}(\boldsymbol{\phi}) = \sum_{i=1}^{m} \boldsymbol{D}_i^t \boldsymbol{\Sigma}_i^{-1} \boldsymbol{D}_i \ (= \mathrm{Cov}(\boldsymbol{s}_m)),$$
$$\boldsymbol{b}(y_i, \boldsymbol{\phi}) = \boldsymbol{U}(\boldsymbol{\phi})^{-1} \Big\{ \boldsymbol{D}_i^t \boldsymbol{\Sigma}_i^{-1} \boldsymbol{g}_i + \boldsymbol{a}_1(\boldsymbol{\phi}) + \frac{1}{2} \boldsymbol{a}_2(\boldsymbol{\phi}) \Big\},$$

where the detailed forms of $\boldsymbol{a}_1(\boldsymbol{\phi})$ and $\boldsymbol{a}_2(\boldsymbol{\phi})$ are given in (4.18) and (4.17) in Section 4.5. It is noted that $\boldsymbol{s}_m = O_p(m^{1/2})$ and $\boldsymbol{U}(\boldsymbol{\phi}) = O(m)$. The following lemma is useful for evaluating the conditional MSE, where the proof is given in Section 4.5.

Lemma 4.1. *Let $\widehat{\boldsymbol{\phi}}_{\rm GT}$ be the solution of estimating equations in* (4.8)*. Then it holds*

$$\begin{aligned} &(\widehat{\boldsymbol{\phi}}_{\rm GT} - \boldsymbol{\phi}) | y_i = \boldsymbol{U}(\boldsymbol{\phi})^{-1} \boldsymbol{s}_m + o_p(m^{-1/2}), \\ &\mathrm{E}[(\widehat{\boldsymbol{\phi}}_{\rm GT} - \boldsymbol{\phi})(\widehat{\boldsymbol{\phi}}_{\rm GT} - \boldsymbol{\phi})^t | y_i] = \boldsymbol{U}(\boldsymbol{\phi})^{-1} + o_p(m^{-1}), \\ &\mathrm{E}[\widehat{\boldsymbol{\phi}}_{\rm GT} - \boldsymbol{\phi} | y_i] = \boldsymbol{b}(y_i, \boldsymbol{\phi}) + o_p(m^{-1}). \end{aligned} \tag{4.11}$$

Lemma 4.1 means that the second-order approximations of the conditional covariance matrix $\mathrm{E}[(\widehat{\boldsymbol{\phi}}_{\rm GT} - \boldsymbol{\phi})(\widehat{\boldsymbol{\phi}}_{\rm GT} - \boldsymbol{\phi})^t | y_i]$ does not depend on y_i, and it coincides with the unconditional results given in Ghosh and Maiti (2004). On the other hand, the second order approximation of the conditional bias $\mathrm{E}[\widehat{\boldsymbol{\phi}}_{\rm GT} - \boldsymbol{\phi} | y_i]$ depends on y_i. It is noted that Lemma 4.1 shows that the estimator $\widehat{\boldsymbol{\phi}}_{\rm GT}$ satisfies Assumption 4.1.

We now derive analytical expressions $T_{2i}(y_i, \boldsymbol{\phi})$ in Theorem 4.1. In the following theorem, we can evaluate $T_{2i}(y_i, \boldsymbol{\phi})$ as

$$T_{2i}(y_i, \boldsymbol{\phi}) = \mathrm{tr}\left[\boldsymbol{P}_i(y_i, \boldsymbol{\phi}) \boldsymbol{U}(\boldsymbol{\phi})^{-1} \right], \tag{4.12}$$

which is $O_p(m^{-1})$, where

$$\boldsymbol{P}_i(y_i, \boldsymbol{\phi}) = (n_i + \nu)^{-2} \begin{pmatrix} \nu^2 Q(m_i)^2 \boldsymbol{x}_i \boldsymbol{x}_i^t & -n_i \nu (n_i + \nu)^{-1} Q(m_i) g_{1i} \boldsymbol{x}_i \\ -n_i \nu (n_i + \nu)^{-1} Q(m_i) g_{1i} \boldsymbol{x}_i^t & n_i^2 (n_i + \nu)^{-2} g_{1i}^2 \end{pmatrix}.$$

Theorem 4.4. *The CMSE of $\widetilde{\mu}_i(y_i, \widehat{\boldsymbol{\phi}}_{\rm GT})$ can be approximated up to $O_p(m^{-1})$ as*

$$\mathrm{CMSE}_i = T_{1i}(y_i, \boldsymbol{\phi}) + T_{2i}(y_i, \boldsymbol{\phi}) + o_p(m^{-1}), \tag{4.13}$$

where $T_{1i}(y_i, \boldsymbol{\phi})$ and $T_{2i}(y_i, \boldsymbol{\phi})$ are given in (4.10) *and* (4.12)*, respectively.*

Proof. From Theorem 4.1, it is sufficient to calculate T_{2i}, which is written as

$$\begin{aligned} \mathrm{E}\Big[\Big\{ (\widehat{\boldsymbol{\phi}}_{\rm GT} - \boldsymbol{\phi})^t \frac{\partial \widetilde{\mu}_i(y_i, \boldsymbol{\phi})}{\partial \boldsymbol{\phi}} \Big\}^2 \Big| y_i \Big] &= \mathrm{trE}\Big[\Big(\frac{\partial \widetilde{\mu}_i}{\partial \boldsymbol{\phi}} \Big) \Big(\frac{\partial \widetilde{\mu}_i}{\partial \boldsymbol{\phi}} \Big)^t (\widehat{\boldsymbol{\phi}}_{\rm GT} - \boldsymbol{\phi})(\widehat{\boldsymbol{\phi}}_{\rm GT} - \boldsymbol{\phi})^t \Big| y_i \Big] \\ &= \mathrm{tr}\Big[\Big(\frac{\partial \widetilde{\mu}_i}{\partial \boldsymbol{\phi}} \Big) \Big(\frac{\partial \widetilde{\mu}_i}{\partial \boldsymbol{\phi}} \Big)^t \mathrm{E}\left[(\widehat{\boldsymbol{\phi}}_{\rm GT} - \boldsymbol{\phi})(\widehat{\boldsymbol{\phi}}_{\rm GT} - \boldsymbol{\phi})^t | y_i \right] \Big]. \end{aligned}$$

It is noted from (4.7) that

$$\frac{\partial \widetilde{\mu}_i(y_i, \boldsymbol{\phi})}{\partial \boldsymbol{\phi}} = \begin{pmatrix} \nu(n_i+\nu)^{-1} Q(m_i)\boldsymbol{x}_i \\ -n_i(n_i+\nu)^{-2} g_{1i} \end{pmatrix}.$$

Then from Lemma 4.1, the last formula can be approximated as

$$\mathrm{tr}\left[\boldsymbol{P}_i(y_i, \boldsymbol{\phi})\boldsymbol{U}(\boldsymbol{\phi})^{-1}\right] + o_p(m^{-1}),$$

which completes the proof. □

Taking the expectation of CMSE_i with respect to y_i, one gets the unconditional MSE given in Theorem 1 of Ghosh and Maiti (2004) with $\delta_i = n_i^{-1}$. In fact,

$$\begin{aligned} T_{1i}(\boldsymbol{\phi}) \equiv & \mathrm{E}[T_{1i}(y_i, \boldsymbol{\phi})] = \frac{\nu}{(n_i+\nu)(\nu - v_2)} Q(m_i), \\ T_{2i}(\boldsymbol{\phi}) \equiv & \mathrm{E}[T_{2i}(y_i, \boldsymbol{\phi})] \\ = & (n_i+\nu)^{-2} \mathrm{tr}\Big[\Big(\begin{matrix} \nu^2 Q(m_i)^2 \boldsymbol{x}_i \boldsymbol{x}_i^t & \boldsymbol{0} \\ \boldsymbol{0}^t & n_i(n_i+\nu)^{-1} Q(m_i)(\nu - v_2)^{-1} \end{matrix} \Big) \boldsymbol{U}(\boldsymbol{\phi})^{-1}\Big]. \end{aligned}$$

Corollary 4.1. *The unconditional MSE of* $\widetilde{\mu}_i(y_i, \widehat{\boldsymbol{\phi}}_{\mathrm{GT}})$ *is approximated as*

$$\mathrm{MSE}_i = T_{1i}(\boldsymbol{\phi}) + T_{2i}(\boldsymbol{\phi}) + o(m^{-1}).$$

It is interesting to investigate the difference between the approximations of the CMSE and the MSE. When the underlying distribution of y_i is a normal distribution, we have $Q(x) = 1$, or $v_0 = 1$ and $v_1 = v_2 = 0$, so that $T_{1i}(y_i, \boldsymbol{\phi}) = 1/(n_i + \nu) = T_{1i}(\boldsymbol{\phi})$, namely the leading term in the CMSE is identical to that in the MSE. Thus, the difference between the CMSE and the MSE appears in the second-order term with $O_p(m^{-1})$. When v_1 or v_2 is not zero, however, the leading term $T_{1i}(y_i, \boldsymbol{\phi})$ in the CMSE is a function of y_i and it is not equal to the leading term $T_{1i}(\boldsymbol{\phi})$ in the MSE. Thus, for distributions far from the normality, the difference between the CMSE and the MSE is significant even when m is large. This tells us about the remark that one cannot replace the conditional MSE given y_i with the corresponding unconditional MSE except for the normal distribution. Some examples including the Poisson and binomial distributions are presented in Section 4.3.3.

We next derive an analytical form of a second-order unbiased estimator for the CMSE. To this end, we define

$$\begin{aligned} \boldsymbol{r}(y_i, \boldsymbol{\phi}) \equiv & \frac{\partial T_{1i}}{\partial \boldsymbol{\phi}} = \begin{pmatrix} \nu(n_i+\nu)^{-1} \lambda_i Q'(\widehat{\mu}_i) Q(m_i)\boldsymbol{x}_i \\ -\lambda_i^2 Q(\widehat{\mu}_i) - \lambda_i n_i (n_i+\nu)^{-2} Q'(\widehat{\mu}_i) g_{1i} \end{pmatrix}, \\ \boldsymbol{R}(y_i, \boldsymbol{\phi}) \equiv & \frac{\partial^2 T_{1i}}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}^t} = \begin{pmatrix} \boldsymbol{T}_{1i}^{11} & \boldsymbol{T}_{1i}^{12} \\ (\boldsymbol{T}_{1i}^{12})^t & T_{1i}^{22} \end{pmatrix}, \end{aligned}$$

where $\lambda_i = (n_i + \nu - v_2)^{-1}$, and

$$\begin{aligned}
\boldsymbol{T}_{1i}^{11} &= (n_i+\nu)^{-2}\nu \boldsymbol{x}_i \boldsymbol{x}_i^t \lambda_i Q(m_i)\left\{2v_2\nu Q(m_i) + Q'(\widetilde{\mu}_i)Q'(m_i)(n_i+\nu)\right\}, \\
\boldsymbol{T}_{1i}^{12} &= \frac{\partial^2 T_{1i}}{\partial\boldsymbol{\beta}\partial\nu} = Q(m_i)\lambda_i(n_i+\nu)^{-2}\left[Q'(\widetilde{\mu}_i)\left\{n_i - \nu(n_i+\nu)\lambda_i\right\} - 2v_2 n_i \nu g_{1i}(n_i+\nu)^{-1}\right]\boldsymbol{x}_i, \\
T_{1i}^{22} &= \frac{\partial^2 T_{1i}}{\partial\nu^2} = 2\lambda_i^3 Q(\widetilde{\mu}_i) + 2\lambda_i^2 n_i (n_i+\nu)^{-2} Q'(\widetilde{\mu}_i) g_{1i} \\
&\qquad + 2\lambda_i n_i (n_i+\nu)^{-4} g_{1i}\left\{(n_i+\nu)Q'(\widehat{\mu}_i) + n_i v_2 g_{1i}\right\}.
\end{aligned}$$

Using (4.11) in Lemma 4.1, we obtain the analytical expressions of T_{11i} and T_{12i} appeared in (4.2) as

$$T_{11i}(y_i,\boldsymbol{\phi}) = \boldsymbol{r}(y_i,\boldsymbol{\phi})^t \boldsymbol{b}(y_i,\boldsymbol{\phi}), \qquad T_{12i}(y_i,\boldsymbol{\phi}) = \frac{1}{2}\mathrm{tr}\left[\boldsymbol{R}(y_i,\boldsymbol{\phi})\boldsymbol{U}(\boldsymbol{\phi})^{-1}\right],$$

thereby the estimator $\widehat{\mathrm{CMSE}}_i$ given in (4.3) is expressed as

$$\widehat{\mathrm{CMSE}}_i = T_{1i}(y_i,\widehat{\boldsymbol{\phi}}_{\mathrm{GT}}) + T_{2i}(y_i,\widehat{\boldsymbol{\phi}}_{\mathrm{GT}}) - \boldsymbol{r}(y_i,\widehat{\boldsymbol{\phi}}_{\mathrm{GT}})^t \boldsymbol{b}(y_i,\widehat{\boldsymbol{\phi}}_{\mathrm{GT}}) - \frac{1}{2}\mathrm{tr}\left[\boldsymbol{R}(y_i,\widehat{\boldsymbol{\phi}}_{\mathrm{GT}})\boldsymbol{U}(\widehat{\boldsymbol{\phi}}_{\mathrm{GT}})^{-1}\right]. \tag{4.14}$$

Theorem 4.5. *The estimator* (4.14) *is a second-order unbiased estimator, namely,*

$$\mathrm{E}[\widehat{\mathrm{CMSE}}_i \mid y_i] = \mathrm{CMSE}_i + o_p(m^{-1}).$$

4.3.3 Some useful examples

We here give representative examples of the NEF-QVF mixed models (4.5) and investigate some properties of the CMSE.

[1] Fay-Herriot model. The Fay-Herriot model suggested in Fay and Herriot (1979) is an area-level model extensively used in small area estimation. The model is described as

$$y_i = \boldsymbol{x}_i^t\boldsymbol{\beta} + v_i + \varepsilon_i, \quad i = 1,\ldots,m,$$

where m is the number of small areas, and v_i's and ε_i's are mutually independently distributed random errors such that $v_i \sim N(0,A)$ and $\varepsilon_i \sim N(0,D_i)$. The notations in (4.5) correspond to $n_i = D_i^{-1}, v_0 = 1, v_1 = v_2 = 0, \mu_i = \theta_i, \nu = A^{-1}$ and $\psi(\theta_i) = \theta_i^2/2$. In this case, the estimating equations in (4.8) reduce to

$$\begin{aligned}
\sum_{i=1}^m (A+D_i)^{-1}\boldsymbol{x}_i y_i &= \sum_{i=1}^m (A+D_i)^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^t\boldsymbol{\beta}, \\
\sum_{i=1}^m (A+D_i)^{-2}(y_i - \boldsymbol{x}_i^t\boldsymbol{\beta})^2 &= \sum_{i=1}^m (A+D_i)^{-1},
\end{aligned}$$

which coincide with the likelihood equations for the maximum likelihood estimators of $\boldsymbol{\beta}$ and A, namely $\widehat{\boldsymbol{\phi}}_{\mathrm{ML}} = \widehat{\boldsymbol{\phi}}_{\mathrm{GT}}$ in Fay-Herriot model. The terms $T_{1i}(y_i,\boldsymbol{\phi})$ and $T_{2i}(y_i,\boldsymbol{\phi})$ in

approximation (8.13) of the CMSE are written as

$$T_{1i}(y_i,\boldsymbol{\phi}) = \frac{AD_i}{A+D_i}$$
$$T_{2i}(y_i,\boldsymbol{\phi}) = \frac{D_i}{(A+D_i)^2}\boldsymbol{x}_i^t\Big(\sum_{j=1}^m \frac{\boldsymbol{x}_j\boldsymbol{x}_j^t}{A+D_j}\Big)^{-1}\boldsymbol{x}_j + \frac{D_i^2(y_i-\boldsymbol{x}_i^t\boldsymbol{\beta})^2}{(A+D_j)^4}\Big(\sum_{j=1}^m \frac{1}{2(A+D_j)^2}\Big)^{-1},$$

which were given in Datta et al. (2011a). In the Fay-Herriot model, $T_{1i}(y_i,\boldsymbol{\phi}) = AD_i/(A+D_i) = T_{1i}(\boldsymbol{\phi})$, namely, the leading terms in the conditional and unconditional MSEs are identical, and the difference is appeared in the term of order $O(m^{-1})$, which is negligible for large m.

[2] Poisson-gamma model. Let $z_1,\ldots,z_m$ be mutually independent random variables having

$$z_i|\lambda_i \sim \mathrm{Po}(n_i\lambda_i) \quad \text{and} \quad \lambda_i \sim \Gamma(\nu m_i, 1/\nu)$$

where $\lambda_1,\ldots,\lambda_m$ are mutually independent, $\mathrm{Po}(\lambda)$ denotes the Poisson distribution with mean λ, and $\Gamma(a,b)$ denotes the gamma distribution with shape parameter a and scale parameter b. Let $y_i = z_i/n_i$ and $\ln m_i = \boldsymbol{x}_i^t\boldsymbol{\beta}$ for $i=1,\ldots,m$. Then, the notations in (4.5) correspond to $v_1=1,\ v_0=v_2=0,\ \mu_i=\lambda_i=\exp(\theta_i)$, and $\psi(\theta_i)=\exp(\theta_i)$. The posterior distribution of λ_i is $\Gamma(\nu m_i+n_iy_i,(n_i+\nu)^{-1})$ or $\Gamma((n_i+\nu)\widehat{\mu}_i,(n_i+\nu)^{-1})$. Then we have

$$T_{1i}(y_i,\boldsymbol{\phi}) = \frac{\widehat{\mu}(y_i,\boldsymbol{\phi})}{n_i+\nu} = \frac{n_iy_i+\nu m_i}{(n_i+\nu)^2},$$

which increases in y_i. Thus, the difference between the conditional and unconditional MSEs increases in y_i. When a large value of y_i is observed, it should be remarked that the conditional MSE of the empirical Bayes estimator given y_i is larger than the unconditional (or integrated) MSE. Hence, it is meaningful to provide practitioners with the information on the conditional MSE as well as the unconditional MSE.

For the Poisson-gamma mixture model, the marginal distribution of y_i (marginal likelihood) is the negative binomial distribution given by

$$f(y_i;\boldsymbol{\phi}) = \frac{\Gamma(n_iy_i+\nu m_i)}{\Gamma(n_iy_i+1)\Gamma(\nu m_i)}\left(\frac{n_i}{n_i+\nu}\right)^{n_iy_i}\left(\frac{\nu}{n_i+\nu}\right)^{\nu m_i},$$

where $\Gamma(\cdot)$ denotes a gamma function. Thus it is noted that the maximum likelihood estimator can be obtained by maximizing $\sum_{i=1}^m \log f(y_i|\boldsymbol{\phi})$.

[3] Binomial-beta model. Let $z_1,\ldots,z_m$ be mutually independent random variables having

$$z_i|p_i \sim \mathrm{Bin}(n_i,p_i) \quad \text{and} \quad p_i \sim B(\nu m_i,\nu(1-m_i)),$$

where $p_1,\ldots,p_m$ are mutually independent, $\mathrm{Bin}(n,p)$ denotes the binomial distribution and $B(a,b)$ denotes the beta distribution. Let $y_i = z_i/n_i$ and $m_i = \exp(\boldsymbol{x}_i^t\boldsymbol{\beta})/\{1+\exp(\boldsymbol{x}_i^t\boldsymbol{\beta})\}$ for $i=1,\ldots,m$. Then the notations in (4.5) correspond to $v_0=0,\ v_1=1$ and $v_2=-1,\ \mu_i = p_i = \exp(\theta_i)/\{1+\exp(\theta_i)\}$ and $\psi(\theta_i)=\ln(1+\exp(\theta_i))$. The posterior distribution of p_i is $B(\nu m_i+n_iy_i, n_i(1-y_i)+\nu(1-m_i))$ or $B((n_i+\nu)\widehat{\mu}_i,(n_i+\nu)(1-\widehat{\mu}_i))$, so that $T_{1i}(y_i,\boldsymbol{\phi})$ is written as

$$T_{1i}(y_i,\boldsymbol{\phi}) = \frac{\widehat{\mu}_i(y_i,\boldsymbol{\phi})\{1-\widehat{\mu}_i(y_i,\boldsymbol{\phi})\}}{n_i+\nu+1},$$

which is a quadratic and concave function of y_i. Since $0 < \widehat{\mu}(y_i, \boldsymbol{\phi}) < 1$, $T_{1i}(y_i, \boldsymbol{\phi})$ is always positive and attains the maximum when $\widehat{\mu}_i = 1/2$ or $y_i = (n_i + \nu)/2n_i - \nu m_i/n_i$, and $T_{1i}(y_i, \boldsymbol{\phi}) = 0$ when $\widehat{\mu}_i = 0$ or 1. Thus, the value of $T_{1i}(y_i, \boldsymbol{\phi})$ is relatively small when y_i is close to 0 or 1. When y_i is around 1/2, the value of $T_{1i}(y_i, \boldsymbol{\phi})$ tends to be larger. When a value around 1/2 is observed for y_i, it should be remarked that the conditional MSE of the EB given y_i is larger than the unconditional (or integrated) MSE.

In the binomial-beta mixture model, the marginal likelihood function is not a familiar form, but proportional to

$$L(\boldsymbol{\phi}) \propto \prod_{i=1}^{m} \frac{B(\nu m_i + n_i y_i, n_i(1-y_i) + \nu(1-m_i))}{B(\nu m_i, \nu(1-m_i))},$$

where $B(\cdot, \cdot)$ is a beta function.

4.4 Numerical Studies

4.4.1 Comparison of CMSE and MSE

We first investigated how different CMSE is from the unconditional MSE. Since the major difference between them appears in the leading terms, namely the terms with order $O_p(1)$ in the CMSE and MSE, we define the ratio of the leading terms in CMSE and MSE as

$$\text{Ratio}_1 = T_{1i}(y_i, \boldsymbol{\phi})/\text{E}[T_{1i}(y_i, \boldsymbol{\phi})],$$

which is a function of y_i and $\boldsymbol{\phi}$. We considered the case $m = 10$, $\nu = 1$, $\boldsymbol{x}_i^t\boldsymbol{\beta} = \mu = 0$ and $n_i = 10$ for $i = 1, \ldots, m$. Then, the curves of the functions Ratio_1 are illustrated Figure 4.1 for the three mixed models: the Fay-Herriot, Poisson-gamma and binomial-beta models. As mentioned before, in the Fay-Herriot model, $\text{Ratio}_1 = 1$ since $T_{1i}(y_i, \boldsymbol{\phi}) = \text{E}\,[T_{1i}(y_i, \boldsymbol{\phi})]$. For the Poisson-gamma and binomial-beta mixture models, Figure 4.1 tells us about the interesting features of their leading terms in the CMSE, namely, the ratio is an increasing function of y_i for the Poisson-gamma mixture model, and a concave and quadratic function of y_i for the binomial-beta mixture model.

We next investigated the corresponding ratios based on the second-order approximations of the CMSE and MSE. Let us define Ratio_2 by

$$\text{Ratio}_2 = \{T_{1i}(y_i, \boldsymbol{\phi}) + T_{2i}(y_i, \boldsymbol{\phi})\}/\text{E}[T_{1i}(y_i, \boldsymbol{\phi}) + T_{2i}(y_i, \boldsymbol{\phi})].$$

Since the second order terms depend on m, we treated three cases $m = 10$, 15 and 20 with $\boldsymbol{x}_i^t\boldsymbol{\beta} = \mu$ and $n_1 = \cdots = n_m = 5$. We used $\widehat{\boldsymbol{\phi}}_{GT}$ for estimation of $\boldsymbol{\phi}$. The performances of Ratio_2 are illustrated in Figure 4.2 for the three mixed models, where the values of (μ, ν) are $(0, 1)$ for the Fay-Herriot model, $(\exp(2), 1)$ for the Poisson-gamma mixture model, and $(\exp(1.5)/\{1 + \exp(1.5)\}, 1)$ for the binomial-beta mixture models. Figure 4.2 demonstrates that the second-order terms for the three mixed models do not contribute so much to Ratio_2 or the conditional MSE.

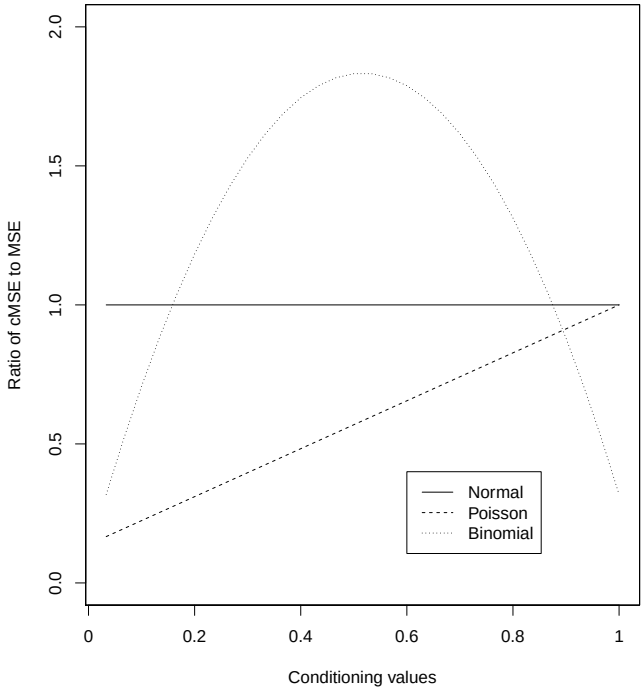

Figure 4.1: Ratio of the leading terms in CMSE and MSE for the Fay-Herriot model, the binomial-beta model, and the Poisson-gamma model.

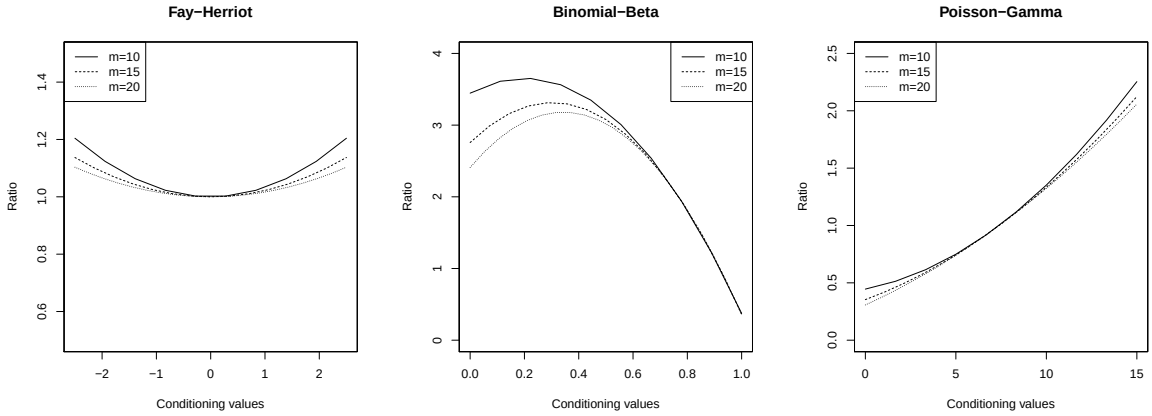

Figure 4.2: Ratio of the leading terms in CMSE and MSE for the Fay-Herriot model, the binomial-beta model, and the Poisson-gamma mixture Model.

4.4.2 Finite performances of the CMSE estimators

We next investigated the finite performances of the second order unbiased estimator of CMSE. We used the Poisson-gamma model and the binomial-beta model with $m = 25$, $n_i = 10$ and

$\nu = 15$. For simplicity, we considered the case without covariates and set $\beta_1 = 0$. Since the conditional MSE depends on the observation, we first obtained the α-quantile point, denoted by $y_{1(\alpha)}$, of the distribution of y_1 for $\alpha = 0.05, 0.25, 0.5, 0.75$ and 0.95. For the Poisson-gamma mixture model, the marginal distribution of y_1 is the negative binomial distribution and $y_{1(\alpha)}$ corresponds to the α-quantile of the negative binomial distribution. For the binomial-beta mixture model, the marginal distribution of y_1 is not a familiar distribution, so that we generated a large number of random samples of y_1 and computed the quantiles.

For computing the true values of CMSE, we generated random samples y_i for $i = 2, \ldots, m$, computed estimates $\widehat{\boldsymbol{\phi}}$ from the simulated data $\{y_{1(\alpha)}, y_2^{(r)}, \ldots, y_m^{(r)}\}$, and calculated the empirical Bayes estimates $\widehat{\mu}_1$ in the 1st area. These procedures were repeated for $R_1 = 10,000$ times to get the true CMSE value in the 1st area under given $y_{1(\alpha)}$:

$$\mathrm{CMSE}_1 = T_{11}(y_{1(\alpha)}, \boldsymbol{\phi}) + \frac{1}{R_1}\sum_{r=1}^{R_1}\Big\{\widehat{\mu}_1^{(r)} - \widetilde{\mu}_1(y_{1(\alpha)}, \boldsymbol{\phi})\Big\}^2,$$

where $\widehat{\mu}_1^{(r)}$ is the empirical Bayes estimates in rth simulation run. For estimating $\boldsymbol{\phi}$, we considered both the GT-estimator and the ML estimator.

Through the same manner as described above, we generate another simulated sample with size $R_2 = 2,000$ and calculate the CMSE estimate $\widehat{\mathrm{CMSE}}_1$. Then, we computed the relative bias (RB) and coefficients of variation (CV) for the CMSE estimator:

$$\mathrm{RB} = \frac{R_2^{-1}\sum_{r=1}^{R_2}\widehat{\mathrm{CMSE}}_1^{(r)} - \mathrm{CMSE}_1}{\mathrm{CMSE}_1},$$

$$\mathrm{CV} = \left\{\frac{1}{R_2}\sum_{r=1}^{R_2}\Big(\widehat{\mathrm{CMSE}}_1^{(r)} - \mathrm{CMSE}_1\Big)^2\right\}^{1/2}\Big/\mathrm{CMSE}_1,$$

where $\widehat{\mathrm{CMSE}}_1^{(r)}$ is the CMSE estimate in the rth replication.

For $\alpha = 0.05$, 0.25, 0.50, 0.75 and 0.95, we report the value of $y_{1(\alpha)}$, CMSE_1, RB and CV in both cases in which we used GT-estimator and ML-estimator for $\boldsymbol{\phi}$, in Table 4.1 for the two mixed models. It is noted that the values of CMSE_1 are multiplied by 100. Table 4.1 demonstrates that the CMSE estimator with the GT-estimator performs well for various values of $y_{1(\alpha)}$ in both models. Concerning the CMSE estimator with the ML-estimator, it is biased than GT, but the $\mathrm{CV}^{\mathrm{ML}}$ is smaller than $\mathrm{CV}^{\mathrm{GT}}$. The true value of CMSE_i has a general trend increasing in $y_{1(\alpha)}$ in the Poisson-gamma model, which coincides with the analytical property discussed the previous section. In the binomial-beta mixture model, the true values of CMSE_i are similar for five α. Altogether, we can conclude that the CMSE estimators with both GT and ML perform well in this setting.

4.4.3 Example: stomach cancer mortality data

We applied the proposed method to the Stomach Cancer Mortality Data and the Infant Mortality Data Before World War II in Japan. The data set consists of the observed number of mortality z_i and its expected number n_i of stomach cancer for women who lived in the ith city or town in Saitama prefecture, Japan, for five years from 1995 to 1999. Such area-level data (z_i, n_i), $i = 1, \ldots, m$, are available for $m = 92$ cities and towns, and the total number of

Table 4.1: Values of CMSE_1, relative bias (RB) and coefficient of variation (CV) of the CMSE estimator for the five conditioning values in the Poisson-gamma and binomial-beta models.

	α	$y_{1(\alpha)}$	$\mathrm{CMSE}_1^{\mathrm{GT}}$	$\mathrm{RB}^{\mathrm{GT}}$	$\mathrm{CV}^{\mathrm{GT}}$	$\mathrm{CMSE}_1^{\mathrm{ML}}$	$\mathrm{RB}^{\mathrm{ML}}$	$\mathrm{CV}^{\mathrm{ML}}$
Poisson-gamma	0.05	0.40	4.10	0.09	0.73	3.92	−0.14	0.20
	0.25	0.70	3.80	0.02	0.53	3.97	−0.30	0.36
	0.50	1.00	4.24	−0.03	0.68	4.31	−0.36	0.41
	0.75	1.30	4.90	0.05	0.71	5.05	−0.30	0.37
	0.95	1.70	6.16	0.06	0.66	6.45	−0.04	0.22
Binomial-beta	0.05	0.10	1.18	−0.10	0.30	1.25	−0.05	0.15
	0.25	0.30	1.07	0.03	0.47	1.10	−0.24	0.30
	0.50	0.40	1.03	0.07	0.56	1.05	−0.32	0.37
	0.75	0.50	1.03	0.06	0.60	1.03	−0.34	0.39
	0.95	0.70	1.06	−0.02	0.51	1.10	−0.23	0.30

mortality in the whole region is $L = 3953$. The expected numbers are adjusted by age on the basis of the population so that $L = \sum_{i=1}^{m} z_i = \sum_{i=1}^{m} n_i$.

For $z_1, \ldots, z_m$, we used the Poisson-gamma model discussed in Section 4.3.3, namely $z_i|\lambda_i \sim \mathrm{Po}(n_i\lambda_i)$ and $\lambda_i \sim \Gamma(\nu m_i, 1/\nu)$. Since data of mortality rate of stomach cancer for men are also available, we can use them as a covariate. Let x_i be a log-transformed mortality rate for men for i-th area. Then, we treat the regression model $\ln m_i = \beta_0 + x_i\beta_1$ for $i = 1, \ldots, m$. The unknown parameters $\boldsymbol{\phi}^t = (\beta_0, \beta_1, \nu)^t$ are estimated as the roots of the estimating equations in (4.8). Their estimates are $\beta_0 = -7.77 \times 10^{-3}$, $\beta_1 = 0.157$ and $\nu = 158$.

To illustrate the difference between CMSE and MSE, we use the percentage relative difference (RD) defined by

$$\mathrm{RD}_i = 100 \times (\widehat{\mathrm{CMSE}}_i - \widehat{\mathrm{MSE}}_i)/\widehat{\mathrm{MSE}}_i.$$

When RD_i is positive, $\widehat{\mathrm{CMSE}}_i$ is larger than $\widehat{\mathrm{MSE}}_i$. In Figure 4.3, the plots of the values $(\widehat{\mathrm{MSE}}_i, \widehat{\mathrm{CMSE}}_i)$ multiplied by $1,000$ and the values of (y_i, RD_i) for $i = 1, \ldots, m$ are given in the left and right figures, respectively, where $y_i = z_i/n_i$ is the standard mortality rate (SMR). From Figure 4.3, it is revealed that the values of $\widehat{\mathrm{CMSE}}_i$ are larger than those of $\widehat{\mathrm{MSE}}_i$ for some areas, and that the relative differences RD_i have great variability, which comes from non-normality of distribution as discussed in Section 4.4.1.

Table 4.2 reports the values of n_i, y_i, EB_i, $\widehat{\mathrm{CMSE}}_i$, $\widehat{\mathrm{MSE}}_i$ and RD_i for ten selected municipalities in Saitama prefecture, where the values of $\widehat{\mathrm{MSE}}_i$ and $\widehat{\mathrm{CMSE}}_i$ are multiplied by $1,000$. It is noted that Kumagaya has the maximum RD value and Yoshida has the minimum RD value in our result. The values of RD tell us about important information that the given empirical Bayes estimate has a different prediction error from the usual unconditional MSE. For instance, in Yoshida, the estimate of the CMSE is 8.631, while that of the unconditional MSE is 18.858, and the resulting RD is -54. This means that the unconditional MSE over-estimates the CMSE. On the other hand, in Kumagaya, the estimate of the CMSE is 7.384, while that of the unconditional MSE is 5.819, and the resulting RD is 27. This means that the unconditional MSE under-estimates the CMSE. Remember that the CMSE is a function

of both y_i and n_i increasing for y_i and decreasing for n_i in the Poisson-gamma model, while the unconditional MSE does not depend on y_i and decreases for n_i. Thus, the CMSE is not always small in areas with small n_i such as Yoshida and Naguri, and the unconditional MSE may over-estimates the CMSE. On the contrary, in area with large n_i such as Kumagaya, the unconditional MSE may under-estimates the CMSE, which leads to a serious situation in real application.

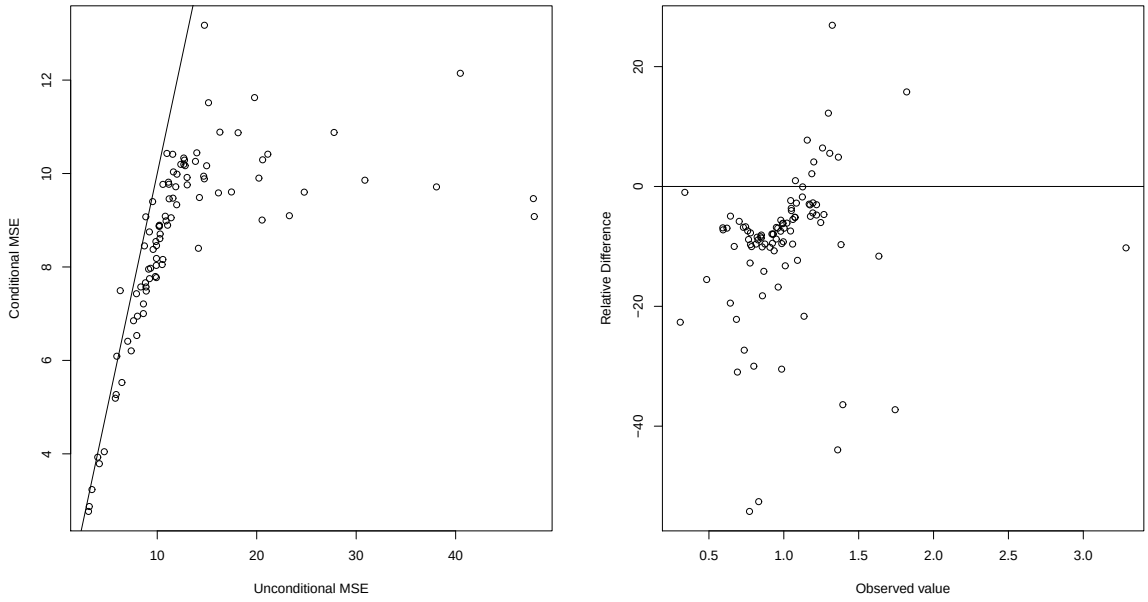

Figure 4.3: Scatter plot of $(\widehat{\mathrm{MSE}}_i, \widehat{\mathrm{CMSE}}_i)$ (left) and that of (y_i, RD_i) (right) for stomach cancer mortality data.

4.4.4 Example: infant mortality data

We next handle the historical data of the Infant Mortality Data Before World War II. The data set consists of the observed number of infant mortality z_i and the number of birth n_i in the i-th city or town in Ishikawa prefecture, Japan, before World War II. Such area-level data are available for $m = 211$ cities, towns and villages, and the total number of infant mortality in the whole region is $L = 4252$. It is noted that the infant mortality rates $y_i = z_i/n_i$ before World War II are not small and distributed around 0.2. Thus, we here apply the data to the binomial-beta model rather than the Poisson-gamma model. For $z_1, \ldots, z_m$, $z_i|p_i$ and p_i have the distributions $z_i|p_i \sim \mathrm{Bin}(n_i, p_i)$ and $p_i \sim B(\nu m_i, \nu(1-m_i))$, where $m_i = \exp(\beta)/(1+\exp(\beta))$ for $i = 1, \ldots, m$, since we do not have any covariates. Thus, the unknown parameters are $\boldsymbol{\phi} = (\beta, \nu)^t$ and their estimates are $\beta = -1.57$, namely $m_i = 0.171$, and $\nu = 102$.

The plots of the values $(\widehat{\mathrm{MSE}}_i, \widehat{\mathrm{CMSE}}_i)$ multiplied by $1,000$ and the values of (y_i, RD_i) for $i = 1, \ldots, m$ are given in the left and right figures of Figure 4.4, respectively. Figure 4.4 suggests that the values of the relative difference RD increases in y_i. This is because the

Table 4.2: Values of expected mortality n_i, SMR y_i, empirical Bayes estimates EB_i, CMSE estimate $\widehat{\mathrm{CMSE}}_i$, unconditional MSE estimate $\widehat{\mathrm{MSE}}_i$ and relative difference RD_i for 10 selected areas in Saitama prefecture.

Area	n_i	y_i	EB_i	$\widehat{\mathrm{CMSE}}_i$	$\widehat{\mathrm{MSE}}_i$	RD_i
Kawagoe	192.1	1.077	1.058	3.892	3.855	1
Kumagaya	102.7	1.324	1.194	7.384	5.819	27
Hatagaya	35.2	1.307	1.114	9.556	9.054	6
Asaka	52.5	1.124	1.031	7.600	7.736	−2
Sakado	51.6	1.298	1.131	8.903	7.933	12
Ooi	20.7	0.867	1.003	9.202	10.720	−14
Naguri	3.6	1.394	0.934	9.435	14.839	−36
Yoshida	6.5	0.771	0.863	8.631	18.858	−54
Kamisato	18.3	1.364	1.066	10.164	9.690	5
Miyashiro	20.1	1.194	1.051	9.516	9.784	−3

leading $O_p(1)$ term is an increasing function of y_i for fixed n_i since y_i is between 0 and 0.5, as investigated in Section 4.4.1. It is observed from Figure 4.4 that the unconditional MSE under-estimates the CMSE in most areas. This gives us a warning message on the empirical Bayes estimates in each area since the unconditional MSE underestimates the estimation error of the empirical Bayes estimate based on given area data. Table 4.3 reports the values of n_i, y_i, EB_i, $\widehat{\mathrm{CMSE}}_i$, $\widehat{\mathrm{MSE}}_i$ and RD_i for fifteen selected municipalities in Ishikawa prefecture, where the values of $\widehat{\mathrm{MSE}}_i$ and $\widehat{\mathrm{CMSE}}_i$ are multiplied by $1,000$. It is noted that Area 175 has the maximum RD value and Area 46 has the minimum RD value in our result. For Area 176, the observed mortality rate $y_i = 0.400$ is much shrunken to $\mathrm{EB}_i = 0.216$ by the empirical Bayes estimator since the number of birth is quite small as given by $n_i = 25$. The unconditional MSE is estimated by 1.216, but the relative difference is $\mathrm{RD}_i = 62$, and the estimate of CMSE is 1.964, which is higher than the MSE estimate. This suggests that it should be good to provide estimates of CMSE as well as estimates of MSE.

4.5 Technical Issues

4.5.1 Proof of Lemma 4.1

For notational simplicity, we put $\boldsymbol{R}_i = \boldsymbol{D}_i^t \boldsymbol{\Sigma}_i^{-1}$ and we use $\boldsymbol{U}$ as $\boldsymbol{U}(\boldsymbol{\phi})$. Using the results in Ghosh and Maiti (2004), we immediately have $\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi} = \boldsymbol{U}^{-1}\boldsymbol{s}_m + o_p(m^{-1/2})$, which implies that

$$\mathrm{E}\left[(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi})(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi})^t | y_i\right] = \boldsymbol{U}^{-1}\mathrm{E}\left[\boldsymbol{s}_m \boldsymbol{s}_m^t | y_i\right]\boldsymbol{U}^{-1} + o_p(m^{-1}),$$

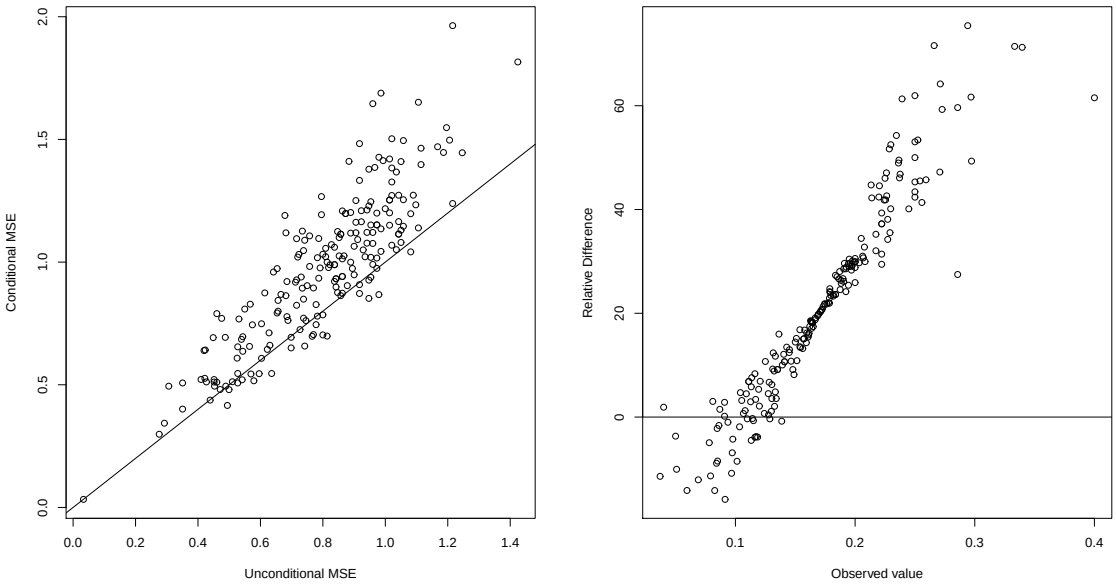

Figure 4.4: Scatter plot of $(\widehat{\mathrm{MSE}}_i, \widehat{\mathrm{CMSE}}_i)$ (left) and that of (y_i, RD_i) (right) for infant mortality data.

Table 4.3: Values of expected mortality n_i, SMR y_i, empirical Bayes estimates EB_i, CMSE estimate $\widehat{\mathrm{CMSE}}_i$, unconditional MSE estimate $\widehat{\mathrm{MSE}}_i$ and relative difference RD_i for 15 selected areas in Ishikawa prefecture.

Area	n_i	y_i	EB_i	$\widehat{\mathrm{CMSE}}_i$	$\widehat{\mathrm{MSE}}_i$	RD_i
1	4146	0.139	0.139	0.033	0.033	0
19	56	0.250	0.199	1.386	0.966	43
23	55	0.164	0.168	1.152	0.973	18
46	197	0.091	0.119	0.416	0.494	−16
71	84	0.060	0.121	0.698	0.814	−14
79	87	0.069	0.124	0.703	0.800	−13
86	101	0.079	0.125	0.658	0.742	−11
96	194	0.119	0.137	0.480	0.499	−4
98	208	0.250	0.224	0.771	0.476	62
112	94	0.160	0.166	0.894	0.770	16
158	173	0.185	0.180	0.685	0.539	27
162	57	0.333	0.229	1.646	0.960	71
175	119	0.294	0.237	1.190	0.678	75
176	25	0.400	0.216	1.964	1.216	62
179	245	0.229	0.212	0.642	0.423	52

where

$$\begin{aligned}\mathrm{E}\left[\boldsymbol{s}_m\boldsymbol{s}_m^t|y_i\right] &= \sum_{j=1}^{m}\mathrm{E}\left[\boldsymbol{R}_j\boldsymbol{g}_j\boldsymbol{g}_j^t\boldsymbol{R}_j^t|y_i\right] = \sum_{j\neq i}^{m}\mathrm{E}\left[\boldsymbol{R}_j\boldsymbol{g}_j\boldsymbol{g}_j^t\boldsymbol{R}_j^t\right] + \boldsymbol{R}_i\boldsymbol{g}_i\boldsymbol{g}_i^t\boldsymbol{R}_i^t \\ &= \boldsymbol{U} + \boldsymbol{R}_i(\boldsymbol{g}_i\boldsymbol{g}_i^t - \boldsymbol{\Sigma}_i)\boldsymbol{R}_i^t,\end{aligned}$$

since $\boldsymbol{g}_j$ depends only on y_j of $\boldsymbol{Y}$ and $y_1,\ldots,y_m$ are mutually independent. Since $\boldsymbol{U} = O(m)$ and $\boldsymbol{R}_i(\boldsymbol{g}_i\boldsymbol{g}_i^t - \boldsymbol{\Sigma}_i)\boldsymbol{R}_i^t = O_p(1)$, we have $\mathrm{E}\left[\boldsymbol{s}_m\boldsymbol{s}_m^t|y_i\right] = \boldsymbol{U} + O_p(1)$, so that

$$\mathrm{E}\left[(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi})(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi})^t|y_i\right] = \boldsymbol{U}^{-1} + o_p(m^{-1}).$$

Next, we evaluate asymptotically the conditional bias of $\widehat{\boldsymbol{\phi}}$, i.e. $\mathrm{E}[\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}|y_i]$. Expanding the equation (4.8) up to second order, we have

$$\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi} = \Big(-\frac{\partial \boldsymbol{s}_m}{\partial \boldsymbol{\phi}}\Big)^{-1}\Big(\boldsymbol{s}_m + \frac{1}{2}\boldsymbol{t} + o_p(1)\Big),$$

where

$$\frac{\partial \boldsymbol{s}_m}{\partial \boldsymbol{\phi}^t} = \sum_{j=1}^{m}\Big(\frac{\partial \boldsymbol{R}_j}{\partial \boldsymbol{\phi}^t}\Big)\left(\boldsymbol{I}_p \otimes \boldsymbol{g}_j\right) + \sum_{j=1}^{m}\boldsymbol{R}_j\Big(\frac{\partial \boldsymbol{g}_j}{\partial \boldsymbol{\phi}^t}\Big),$$

noting that $\partial \boldsymbol{s}_m/\partial \boldsymbol{\phi}^t = -\boldsymbol{U} + o_p(m)$, and

$$\boldsymbol{t} = \mathbf{col}_\ell\Big\{(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi})^t\left(\frac{\partial^2 S_{m\ell}}{\partial \boldsymbol{\phi}\partial \boldsymbol{\phi}^t}\right)(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi})\Big\},$$

for $\boldsymbol{s}_m = (S_{m1},\ldots,S_{mq})$ with $q = p+1$. It noted that $S_{mk} = \boldsymbol{R}_{ik}\boldsymbol{g}_i$ for $k = 1,\ldots,q$, where $\boldsymbol{R}_{ik}$ is the k-th row vector of $\boldsymbol{R}_i$. The notation $\mathbf{col}_\ell\{a_\ell\}$ for scalars a_ℓ's, $\ell = 1,\ldots,n$ is defined by

$$\mathbf{col}_\ell\{a_\ell\} = (a_1, a_2, \ldots, a_n)^t.$$

Let $\boldsymbol{W} = \partial \boldsymbol{s}_m/\partial \boldsymbol{\phi}^t - (-\boldsymbol{U})$, then we have

$$\Big(-\frac{\partial \boldsymbol{s}_m}{\partial \boldsymbol{\phi}}\Big)^{-1} = -\boldsymbol{U}^{-1} - \boldsymbol{U}^{-1}\boldsymbol{W}\boldsymbol{U}^{-1} + o_p(m^{-3/2}).$$

Therefore, it follows that

$$\begin{aligned}\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi} &= \Big\{\boldsymbol{U}^{-1} + \boldsymbol{U}^{-1}\boldsymbol{W}\boldsymbol{U}^{-1} + o_p(m^{-3/2})\Big\}\Big\{\boldsymbol{s}_m + \frac{1}{2}\boldsymbol{t} + o_p(1)\Big\} \\ &= \boldsymbol{U}^{-1}\boldsymbol{s}_m + \frac{1}{2}\boldsymbol{U}^{-1}\boldsymbol{t} + \boldsymbol{U}^{-1}\boldsymbol{W}\boldsymbol{U}^{-1}\boldsymbol{s}_m + o_p(m^{-1}),\end{aligned}$$

whereby

$$\mathrm{E}[\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}|y_i] = \boldsymbol{U}^{-1}\boldsymbol{R}_i\boldsymbol{g}_i + \frac{1}{2}\boldsymbol{U}^{-1}\mathrm{E}\left[\boldsymbol{t}|y_i\right] + \boldsymbol{U}^{-1}\mathrm{E}\left[\boldsymbol{W}\boldsymbol{U}^{-1}\boldsymbol{s}_m|y_i\right]. \tag{4.15}$$

For the second term in (4.15), note that

$$\begin{aligned}\mathrm{E}\,[\boldsymbol{t}|y_i] &= \mathbf{col}_\ell\Big\{E\left[(\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi})^t\left(\frac{\partial^2 S_{m\ell}}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}^t}\right)(\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi})\middle|y_i\right]\Big\}\\ &= \mathbf{col}_\ell\Big\{\mathrm{tr}\Big\{\left(\frac{\partial^2 S_{m\ell}}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}^t}\right)\mathrm{E}\left[(\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi})^t(\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi})\middle|y_i\right]\Big\}\Big\}\\ &= \mathbf{col}_\ell\Big\{\mathrm{tr}\Big(\mathrm{E}\Big[\frac{\partial^2 S_{m\ell}}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}^t}\Big]\boldsymbol{U}^{-1}\Big)\Big\}+o_p(1)\equiv\boldsymbol{a}_2(\boldsymbol{\phi})+o_p(1).\end{aligned}$$

The straightforward calculation shows that

$$\frac{\partial^2 S_{m\ell}}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}^t}=\sum_{i=1}^m\left\{\left(\frac{\partial^2\boldsymbol{R}_{i\ell}}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}^t}\right)(\boldsymbol{I}_p\otimes\boldsymbol{g}_i)+2\frac{\partial\boldsymbol{R}_{i\ell}}{\partial\boldsymbol{\phi}}\frac{\partial\boldsymbol{g}_i}{\partial\boldsymbol{\phi}^t}+(\boldsymbol{I}_p\otimes\boldsymbol{R}_{i\ell})\left(\frac{\partial^2\boldsymbol{g}_i}{\partial\boldsymbol{\phi}^t\partial\boldsymbol{\phi}}\right)\right\},$$

so that

$$\mathrm{E}\left[\frac{\partial^2 S_{m\ell}}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}^t}\right]=\sum_{i=1}^m\left\{2\left(\frac{\partial\boldsymbol{R}_{i\ell}}{\partial\boldsymbol{\phi}}\right)\boldsymbol{D}_i+(\boldsymbol{I}_p\otimes\boldsymbol{R}_{i\ell})\mathrm{E}\left(\frac{\partial^2\boldsymbol{g}_i}{\partial\boldsymbol{\phi}^t\partial\boldsymbol{\phi}}\right)\right\}.$$

Since

$$\frac{\partial\boldsymbol{g}_i}{\partial\boldsymbol{\phi}^t}=2Q(m_i)(y_i-m_i)\begin{pmatrix}\mathbf{0}^t & 0\\ \boldsymbol{x}_i^t & 0\end{pmatrix}-\boldsymbol{D}_i, \tag{4.16}$$

we obtain

$$\frac{\partial^2\boldsymbol{g}_i}{\partial\boldsymbol{\phi}^t\partial\boldsymbol{\phi}}=\begin{pmatrix}2\boldsymbol{x}_iQ(m_i)\left\{Q'(m_i)(y_i-m_i)-Q(m_i)\right\}\\ 0\end{pmatrix}\otimes\begin{pmatrix}\mathbf{0}^t & 0\\ \boldsymbol{x}_i^t & 0\end{pmatrix}-\frac{\partial\boldsymbol{D}_i}{\partial\boldsymbol{\phi}},$$

whereby

$$\boldsymbol{Z}_i\equiv E\left(\frac{\partial^2\boldsymbol{g}_i}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}^t}\right)=\begin{pmatrix}-2\boldsymbol{x}_iQ(m_i)^2\\ 0\end{pmatrix}\otimes\begin{pmatrix}\mathbf{0} & \boldsymbol{x}_i\\ 0 & 0\end{pmatrix}-\frac{\partial\boldsymbol{D}_i}{\partial\boldsymbol{\phi}}.$$

Then we have

$$\boldsymbol{a}_2(\boldsymbol{\phi})=\mathbf{col}_\ell\Big\{\mathrm{tr}\Big(\boldsymbol{U}^{-1}\sum_{i=1}^m\Big\{2\Big(\frac{\partial\boldsymbol{R}_{i\ell}}{\partial\boldsymbol{\phi}}\Big)\boldsymbol{D}_i+(\boldsymbol{I}_q\otimes\boldsymbol{R}_{i\ell})\boldsymbol{Z}_i\Big\}\Big)\Big\}. \tag{4.17}$$

For the evaluation of the third term in (4.15), we get

$$\boldsymbol{U}^{-1}E\left[\boldsymbol{W}\boldsymbol{U}^{-1}\boldsymbol{s}_m|y_i\right]=\boldsymbol{U}^{-1}E\left[\boldsymbol{W}\boldsymbol{U}^{-1}\boldsymbol{s}_m\right]+o_p(m^{-1}),$$

and

$$\begin{aligned}\mathrm{E}\left[\boldsymbol{W}\boldsymbol{U}^{-1}\boldsymbol{s}_m\right] &= \mathrm{E}\left[\left(\frac{\partial\boldsymbol{s}_m}{\partial\boldsymbol{\phi}^t}\right)\boldsymbol{U}^{-1}\boldsymbol{s}_m\right]\\ &= \sum_{i=1}^m\left(\frac{\partial\boldsymbol{R}_i}{\partial\boldsymbol{\phi}^t}\right)\mathrm{E}\left[(\boldsymbol{I}_p\otimes\boldsymbol{g}_i)\boldsymbol{U}^{-1}\boldsymbol{R}_i\boldsymbol{g}_i\right]+\sum_{i=1}^m\boldsymbol{R}_i\mathrm{E}\left[\left(\frac{\partial\boldsymbol{g}_i}{\partial\boldsymbol{\phi}^t}\right)\boldsymbol{U}^{-1}\boldsymbol{R}_i\boldsymbol{g}_i\right].\end{aligned}$$

Using the expression (4.16), we finally have

$$\begin{aligned}\boldsymbol{a}_1(\boldsymbol{\phi}) &\equiv \mathrm{E}\left[\boldsymbol{W}\boldsymbol{U}^{-1}\boldsymbol{s}_m\right]\\ &= \sum_{i=1}^m\left(\frac{\partial\boldsymbol{R}_i}{\partial\boldsymbol{\phi}^t}\right)\mathrm{vec}(\boldsymbol{D}_i\boldsymbol{U}^{-1})+2\sum_{i=1}^m\boldsymbol{R}_iQ(m_i)\begin{pmatrix}\mathbf{0}^t & 0\\ \boldsymbol{x}_i^t & 0\end{pmatrix}\boldsymbol{U}^{-1}\boldsymbol{R}_i\begin{pmatrix}\mu_{2i}\\ \mu_{3i}\end{pmatrix},\end{aligned} \tag{4.18}$$

which completes the proof. □

4.5.2 Numerical evaluation of partial derivatives.

The analytical expression of $\partial \boldsymbol{R}_i/\partial \boldsymbol{\phi}^t$ and $\partial \boldsymbol{D}_i/\partial \boldsymbol{\phi}$ are complex and not practical. However, the values of these derivatives at some value $\boldsymbol{\phi}_0$ can be easily calculated. Let z_m be a positive number depending on m, then the value of $\partial \boldsymbol{R}_i/\partial \phi_k,\ k = 1, \ldots, k$ at $\boldsymbol{\phi} = \boldsymbol{\phi}_0$ is evaluated as

$$\frac{\partial \boldsymbol{R}_i}{\partial \phi_k}(\boldsymbol{\phi}_0^*) \equiv (2z_m)^{-1} \left\{ \boldsymbol{R}_i(\boldsymbol{\phi}_0 + z_m \boldsymbol{e}_k) - \boldsymbol{R}_i(\boldsymbol{\phi}_0 - z_m \boldsymbol{e}_k) \right\},$$

where $\boldsymbol{e}_k$ is a vector of 0's other than k-th element is 1. Since the difference between $\partial \boldsymbol{R}_i/\partial \phi_k$ and $\partial \boldsymbol{R}_i/\partial \phi_k^*$ at $\boldsymbol{\phi} = \boldsymbol{\phi}_0$ is $O(z_m)$, the choice $z_m = o(m^{-1})$ does not affect the second-order unbiasedness of the CMSE estimator established in Theorem 4.5. In numerical studies given in this paper, we choose $z_m = m^{-5/4}$ satisfying $z_m = o(m^{-1})$. The partial derivative $\partial \boldsymbol{D}_i/\partial \boldsymbol{\phi}$ can be numerically evaluated in the same way.

Chapter 5

Heteroscedastic Nested Error Regression Models

5.1 Introduction

5.1.1 Nested error regression model

In some applications, unit level data can be available. For such a case, Battese et al. (1988) suggested the nested error regression (NER) model described as

$$y_{ij} = \boldsymbol{x}_{ij}^t \boldsymbol{\beta} + v_i + \varepsilon_{ij}, \quad i = 1, \ldots, m, \quad j = 1, \ldots, n_i, \tag{5.1}$$

where y_{ij} is the observed response value, $\boldsymbol{x}_{ij}$ is a vector of associated covariates, n_i is the area sample size which is typically small, v_i and ε_{ij} are the random effect and sampling error. Here it is assumed that v_i and ε_{ij} are mutually independent and they hold $\mathrm{E}[v_i] = \mathrm{E}[\varepsilon_{ij}] = 0$, $\mathrm{Var}(v_i) = \tau^2$ and $\mathrm{Var}(\varepsilon_{ij}) = \sigma^2$, noting that the normality is often added for these variables. The model (5.1) can be regarded as the random intercept model in the general linear mixed model, and the model (5.1) is also useful in biological experiments and econometric analysis. The typical purpose in (5.1) is the estimating (predicting) area-specific quantity $\mu_i = \boldsymbol{c}_i^t \boldsymbol{\beta} + v_i$, and it is well-known that the best linear predictor (BLP) has the form

$$\widetilde{\mu}_i \equiv \widetilde{\mu}_i(\boldsymbol{y}_i, \boldsymbol{\phi}) = \boldsymbol{c}_i^t \boldsymbol{\beta} + \frac{n_i \tau^2}{n_i \tau^2 + \sigma^2} (\bar{y}_i - \bar{\boldsymbol{x}}_i^t \boldsymbol{\beta}), \tag{5.2}$$

where $\bar{y}_i = n_i^{-1} \sum_{j=1}^{n_i} y_{ij}$ and $\bar{\boldsymbol{x}}_i = n_i^{-1} \sum_{j=1}^{n_i} \boldsymbol{x}_{ij}$ are the sample means in the ith area, and $\boldsymbol{\phi} = (\boldsymbol{\beta}^t, \tau^2, \sigma^2)^t$ is the model parameters in (5.1). To use the BLP in practice, we need to estimate the unknown parameter $\boldsymbol{\phi}$ from the data, and several estimator including maximum likelihood estimator and the moment estimator have been suggested. Then the empirical best linear unbiased predictor (EBLUP) is obtained as $\widehat{\mu}_i = \widetilde{\mu}_i(\boldsymbol{y}_i, \widehat{\boldsymbol{\phi}})$. Typically, it holds $\widehat{\mu}_i - \widetilde{\mu}_i \to 0$ as $m \to \infty$ if $\widehat{\boldsymbol{\phi}}$ is a consistent estimator of $\boldsymbol{\phi}$.

From (5.1), it follows that $\mathrm{Var}(y_{ij}) = \tau^2 + \sigma^2$, which means that the variances of the observations are equal over all areas. Though Battese et al. (1988) applied the NER model to crop data in Iowa counties, Jiang and Nguyen (2012) illustrated that the within-area sample variances change dramatically from small-area to small-area for the data. This motivate us to extend the traditional NER model to cases with heteroscedastic variances.

5.1.2 Unstructured heteroscedastic variances

Jiang and Nguyen (2012) proposed the heteroscedastic nested error regression (HNER) model by assuming $v_i \sim N(0, \lambda\sigma_i^2)$ and $\varepsilon_{ij} \sim N(0, \sigma_i^2)$ in (5.1). In their model, the number of parameters diverges as $m \to \infty$, namely Neyman-Scott problem is occurred. However, they showed that the maximum likelihood estimators of λ and $\boldsymbol{\beta}$ obtained as the minimizer of the negative profile log-likelihood function without irrelevant constants $Q(\boldsymbol{\beta}, \lambda)$ given by

$$Q(\boldsymbol{\beta}, \lambda) = \sum_{i=1}^{m} \left\{ n_i \log(s_i^2) + \log(1 + n_i\lambda) \right\},$$

with

$$s_i^2 = \frac{1}{n_i} \left\{ \sum_{j=1}^{n_i} (y_{ij} - \boldsymbol{x}_{ij}^t\boldsymbol{\beta})^2 - \frac{\lambda}{1 + n_i\lambda} (\bar{y}_i - \bar{\boldsymbol{x}}_i^t\boldsymbol{\beta})^2 \right\},$$

are consistent as $m \to \infty$ although heteroscedastic variances σ_i^2 are inconsistent. They also pointed out that BLP given in (5.2) can be rewritten as

$$\widetilde{\mu}_i = \boldsymbol{c}_i^t\boldsymbol{\beta} + \frac{n_i\lambda}{1 + n_i\lambda} (\bar{y}_i - \bar{\boldsymbol{x}}_i^t\boldsymbol{\beta}),$$

which are free from σ_i^2. Therefore, EBLUP is asymptotically equivalent to BLP as $m \to \infty$ even though σ_i^2 is inconsistent.

For measuring uncertainty of EBLUP, the mean squared errors (MSE) are often used in small area estimation, which is defined as $\mathrm{MSE}_i = \mathrm{E}\left[(\widehat{\mu}_i - \mu_i)^2\right]$. In the model by Jiang and Nguyen (2012), it was shown that

$$\mathrm{E}\left[(\widetilde{\mu}_i - \mu_i)^2\right] = \frac{\lambda\sigma_i^2}{1 + n_i\lambda},$$

Which depends on σ_i^2. Hence, the asymptotically valid MSE estimator cannot be obtained under unstructured heteroscedastic variances.

5.1.3 Random heteroscedastic variances

To overcome the inconsistency of the estimating heteroscedastic variances σ_i^2, Kubokawa et al. (2016) suggested the following hierarchical random dispersion structure:

$$v_i \sim N(0, \lambda\sigma_i^2), \quad \varepsilon_{ij} \sim N(0, \sigma_i^2), \quad \sigma_i^{-2} \sim \Gamma(\tau_1/2, \tau_2/2),$$

where τ_1 and τ_2 are unknown parameters to characterize the randomness of heteroscedastic variances σ_i^2. Therefore the model parameters are $\boldsymbol{\beta}, \lambda$ and two dispersion parameter τ_1 and τ_2. Under the setting, they showed that BLP of μ_i is identical to that of Jiang and Nguyen (2012). Due to the conjugacy of the inverse gamma distribution, the marginal distribution of $\boldsymbol{y}_i$ is obtained in the closed form. Hence, they considered the maximum likelihood estimation by maximizing the following function:

$$\begin{aligned} Q(\boldsymbol{\beta}, \lambda, \tau_1, \tau_2) = m\tau_1 \log \tau_2 + 2\sum_{i=1}^{m} \log \Gamma\left(\frac{n_i + \tau_1}{2}\right) - 2m \log \Gamma\left(\frac{\tau_1}{2}\right) \\ - \sum_{i=1}^{m} \log(1 + n_i\lambda) - \sum_{i=1}^{m} (n_i + \tau_1) \log(R_i + \tau_2), \end{aligned}$$

where

$$R_i = \sum_{j=1}^{n_i}(y_{ij} - \boldsymbol{x}_{ij}^t\boldsymbol{\beta})^2 - \frac{n_i^2\lambda}{1+n_i\lambda}(\bar{y}_i - \bar{\boldsymbol{x}}_i^t\boldsymbol{\beta})^2.$$

They also proved that the maximum likelihood estimators are consistent, so that EBLUP is asymptotically valid and consistent MSE estimator can be constructed. However, in simulation studies, it has been revealed that the finite sample performances of the estimator of τ_1, τ_2 tend to be unstable.

5.1.4 Heteroscedastic variances with variance functions

While these two heteroscedastic variance models are useful, the serious drawback of the two models is that both require normality assumption for random effects and error terms, which are not necessary satisfied in real application. Hence, the purpose of this paper is to address the issue of relaxing assumptions of classical normal NER models toward two directions: heteroscedasticity of variances and non-normality of underlying distributions.

In real data analysis, we often encounter the situation where the sampling variance $\mathrm{Var}(\varepsilon_{ij})$ is affected by the covariate $\boldsymbol{x}_{ij}$. In such case, the variance function is a useful tool for describing its relationship. Variance function estimation has been studied in the literature in the framework of heteroscedastic nonparametric regression. For example, see Hall and Carroll (1989), Muller and Stadtmuller (1987) and Ruppert et al. (1997). Thus, in this paper, we propose the use of the technique to introduce the heteroscedastic variances into the NER model without assuming normality of underlying distributions. The variance structure we consider is $\mathrm{Var}(y_{ij}) = \tau^2 + \sigma_{ij}^2$, namely, the setup means that the sampling error ε_{ij} has heteroscedastic variance $\mathrm{Var}(\varepsilon_{ij}) = \sigma_{ij}^2$. Then we suggest the variance function model given by $\sigma_{ij}^2 = \sigma^2(\boldsymbol{z}_{ij}^t\boldsymbol{\gamma})$, where the details are explained in Section 5.2. In terms of modeling the heteroscedastic variances with covariates, the generalized linear mixed models (Jiang, 2006) are also the useful tool. The small area models using generalized linear mixed models are investigated in Ghosh et al. (1998). However, the generalized linear mixed model requires strong parametric assumption compared to the heteroscedastic model without assuming underlying distributions proposed in this paper. Hence, the generalized linear mixed model seems still restrictive while it is an attractive method for modeling heteroscedasticity in variances.

In this paper, we propose flexible and tractable HNER models without assuming normality for either v_i nor ε_{ij}. The advantage of the proposed model is that the MSE of the EB or EBLUP and its unbiased estimator are derived analytically in closed forms up to second-order without assuming normality for v_i and ε_{ij}. Most estimators of the MSE have been given by numerical methods such as Jackknife and bootstrap methods except for Lahiri and Rao (1995), who provided an analytical second-order unbiased estimator of the MSE in the Fay-Heriot model. Hall and Maiti (2006b) developed a moment matching bootstrap method for nonparametric estimation of MSE in nested error regression models. The suggested method is actually convenient but it requires bootstrap replication and has computational burden. In this paper, without assuming the normality, we derive a closed expression for a second-order unbiased estimator of the MSE using second-order biases and variances of estimators of the model parameters. Thus our MSE estimator does not require any resampling method and is convenient in practical use. Also our MSE estimator can be regarded as a generalization of the robust MSE estimator given in Lahiri and Rao (1995).

In Section 5.2, we describe the proposed HNER model with variance functions, and provide the moment method for estimating model parameters without assuming normality for both random effects and error terms. We also derive some asymptotic properties of the proposed estimator. In Section 5.3, we consider the problem of predicting μ_i, and derive BLP and EBLUP. Moreover, a second order unbiased estimator of MSE is constructed in the analytical way. We then present some numerical studies and an application to real data set in Section 5.4. All technical proofs are given in Section 5.5.

5.2 HNER Models with Variance Functions

5.2.1 Model settings

Suppose that there are m small clusters, and let $(y_{i1}, \boldsymbol{x}_{i1}), \ldots, (y_{in_i}, \boldsymbol{x}_{in_i})$ be the pairs of n_i observations from the i-th cluster, where $\boldsymbol{x}_{ij}$ is a p-dimensional known vector of covariates. We consider the heteroscedastic nested error regression model

$$y_{ij} = \boldsymbol{x}_{ij}^t\boldsymbol{\beta} + v_i + \varepsilon_{ij}, \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, m, \tag{5.3}$$

where $\boldsymbol{\beta}$ is a p-dimensional unknown vector of regression coefficients, and v_i and ε_{ij} are mutually independent random variables with mean zero and variances $\mathrm{Var}(v_i) = \tau^2$ and $\mathrm{Var}(\varepsilon_{ij}) = \sigma_{ij}^2$, which are denoted by

$$v_i \sim (0, \tau^2) \quad \text{and} \quad \varepsilon_{ij} \sim (0, \sigma_{ij}^2). \tag{5.4}$$

It is noted that no specific distributions are assumed for v_i and ε_{ij}. It is assumed that the heteroscedastic variance σ_{ij}^2 of ε_{ij} is given by

$$\sigma_{ij}^2 = \sigma^2(\boldsymbol{z}_{ij}^t\boldsymbol{\gamma}), \quad i = 1, \ldots, m, \tag{5.5}$$

where $\boldsymbol{z}_{ij}$ is a q-dimensional known vector given for each cluster, and $\boldsymbol{\gamma}$ is a q-dimensional unknown vector. The variance function $\sigma^2(\cdot)$ is a known (user specified) function whose range is nonnegative. Some examples of the variance function are given below. The model parameters are $\boldsymbol{\beta}$, τ^2 and $\boldsymbol{\gamma}$, and the total number of the model parameters is $p + q + 1$.

Let $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})^t$, $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in_i})^t$ and $\boldsymbol{\epsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{in_i})^t$. Then the model (5.3) is expressed in a vector form as

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

where $\mathbf{1}_n$ is an $n \times 1$ vector with all elements equal to one, and the covariance matrix of $\boldsymbol{\epsilon}_i$ is

$$\boldsymbol{\Sigma}_i = \mathrm{Var}(\boldsymbol{y}_i) = \tau^2\boldsymbol{J}_{n_i} + \boldsymbol{W}_i,$$

for $\boldsymbol{J}_{n_i} = \mathbf{1}_{n_i}\mathbf{1}_{n_i}'$ and $\boldsymbol{W}_i = \mathrm{diag}(\sigma_{i1}^2, \ldots, \sigma_{in_i}^2)$. It is noted that the inverse of $\boldsymbol{\Sigma}_i$ is expressed as

$$\boldsymbol{\Sigma}_i^{-1} = \boldsymbol{W}_i^{-1}\left(\boldsymbol{I}_{n_i} - \frac{\tau^2\boldsymbol{J}_{n_i}\boldsymbol{W}_i^{-1}}{1 + \tau^2\sum_{j=1}^{n_i}\sigma_{ij}^{-2}}\right),$$

where $\boldsymbol{W}_i^{-1} = \mathrm{diag}(\sigma_{i1}^{-2}, \ldots, \sigma_{in_i}^{-2})$. Further, let $\boldsymbol{y} = (\boldsymbol{y}_1^t, \ldots, \boldsymbol{y}_m^t)^t$, $\boldsymbol{X} = (\boldsymbol{X}_1^t, \ldots, \boldsymbol{X}_m^t)^t$, $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1^t, \ldots, \boldsymbol{\epsilon}_m^t)^t$ and $\boldsymbol{v} = (v_1\mathbf{1}_{n_1}^t, \ldots, v_m\mathbf{1}_{n_m}^t)^t$. Then, the matricidal form of (5.3) is written as $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{v} + \boldsymbol{\epsilon}$, where $\mathrm{Var}(\boldsymbol{y}) = \boldsymbol{\Sigma} = \text{block diag}(\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_m)$. Now we give three examples of the variance function $\sigma^2(\boldsymbol{z}_{ij}^t\boldsymbol{\gamma})$ in (5.5).

(a) In the case that the dispersion of the sampling error is proportional to the mean, it is reasonable to put $\boldsymbol{z}_{ij} = \boldsymbol{x}_{(s)ij}$ and $\sigma^2(\boldsymbol{x}_{(s)ij}^t\boldsymbol{\gamma}) = (\boldsymbol{x}_{(s)ij}^t\boldsymbol{\gamma})^2$ for a sub-vector $\boldsymbol{x}_{(s)ij}$ of the covariate $\boldsymbol{x}_{ij}$. For identifiability of $\boldsymbol{\gamma}$, we restrict $\gamma_1 > 0$.

(b) Consider the case that m clusters are decomposed into q homogeneous groups $S_1, \ldots, S_q$ with $\{1, \ldots, m\} = S_1 \cup \ldots \cup S_q$. Then, we put

$$\boldsymbol{z}_{ij} = \left(1_{\{i \in S_1\}}, \ldots, 1_{\{i \in S_q\}}\right)^t,$$

which implies that

$$\sigma_{ij}^2 = \gamma_t^2 \quad \text{for} \quad i \in S_t.$$

Note that $\mathrm{Var}(y_{ij}) = \tau^2 + \gamma_t^2$ for $i \in S_t$. Thus, the models assumes that the m clusters are divided into known q groups with their variance are equal over the same groups. Jiang and Nguyen (2012) used a similar setting and argued that the unbiased estimator of the heteroscedastic variance is consistent when $|S_k| \to \infty, k = 1, \ldots, q$ as $m \to \infty$, where $|S_k|$ denotes the number of elements in S_k.

(c) Log linear functions of variance were treated in Cook and Weisberg (1983) and others. That is, $\log \sigma_{ij}^2$ is a linear function, and σ_{ij}^2 is written as $\sigma^2(\boldsymbol{z}_{ij}^t\boldsymbol{\gamma}) = \exp(\boldsymbol{z}_{ij}^t\boldsymbol{\gamma})$. Similarly to (a), we put $\boldsymbol{z}_{ij} = \boldsymbol{x}_{(s)ij}$.

For the above two cases (a) and (b), we have $\sigma^2(x) = x^2$, while the case (c) corresponds to $\log\{\sigma^2(x)\} = x$. In simulation and empirical studies in Section 8.4, we use the log-linear variance model. As given in the subsequent section, we show consistency and asymptotic expression of estimators for $\boldsymbol{\gamma}$ as well as $\boldsymbol{\beta}$ and τ^2.

5.2.2 Estimation

We here provide estimators of the model parameters $\boldsymbol{\beta}$, τ^2 and $\boldsymbol{\gamma}$. When values of $\boldsymbol{\gamma}$ and τ^2 are given, the vector $\boldsymbol{\beta}$ of regression coefficients is estimated by the generalized least squares (GLS) estimator

$$\widetilde{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\beta}}(\tau^2, \boldsymbol{\gamma}) = (\boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{y} = \left(\sum_{i=1}^m \boldsymbol{X}_i^t\boldsymbol{\Sigma}_i^{-1}\boldsymbol{X}_i\right)^{-1} \sum_{i=1}^m \boldsymbol{X}_i^t\boldsymbol{\Sigma}_i^{-1}\boldsymbol{y}_i. \tag{5.6}$$

This is not a feasible form since $\boldsymbol{\gamma}$ and τ^2 are unknown. When estimators $\widehat{\tau}^2$ and $\widehat{\boldsymbol{\gamma}}$ are used for τ^2 and $\boldsymbol{\gamma}$, we get the feasible estimator $\widehat{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\beta}}(\widehat{\tau}^2, \widehat{\boldsymbol{\gamma}})$ by replacing τ^2 and $\boldsymbol{\gamma}$ in $\widetilde{\boldsymbol{\beta}}$ with their estimators.

Concerning estimation of τ^2, we use the second moment of observations y_{ij}'s. From model (5.3), it is seen that

$$E\left[(y_{ij} - \boldsymbol{x}_{ij}^t\boldsymbol{\beta})^2\right] = \tau^2 + \sigma^2(\boldsymbol{z}_{ij}^t\boldsymbol{\gamma}). \tag{5.7}$$

Based on the ordinary least squares (OLS) estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y}$, a moment estimator of τ^2 is given by

$$\widehat{\tau}^2 = \frac{1}{N}\sum_{i=1}^m \sum_{j=1}^{n_i} \left\{(y_{ij} - \boldsymbol{x}_{ij}^t\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}})^2 - \sigma^2(\boldsymbol{z}_{ij}^t\boldsymbol{\gamma})\right\}, \tag{5.8}$$

with substituting estimator $\widehat{\boldsymbol{\gamma}}$ into $\boldsymbol{\gamma}$, where $N = \sum_{i=1}^{m} n_i$.

For estimation of $\boldsymbol{\gamma}$, we consider the within difference in each cluster. Let $\bar{y}_i$ be the sample mean in the i-th cluster, namely $\bar{y}_i = n_i^{-1} \sum_{j=1}^{n_i} y_{ij}$. It is noted that for $\bar{\varepsilon}_i = n_i^{-1} \sum_{j=1}^{n_i} \varepsilon_{ij}$,

$$y_{ij} - \bar{y}_i = (\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)^t \boldsymbol{\beta} + (\varepsilon_{ij} - \bar{\varepsilon}_i),$$

which dose not include the term of v_i. Then it is seen that

$$E\left[\left\{y_{ij} - \bar{y}_i - (\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)^t \boldsymbol{\beta}\right\}^2\right] = \left(1 - 2n_i^{-1}\right) \sigma^2(\boldsymbol{z}_{ij}^t \boldsymbol{\gamma}) + n_i^{-2} \sum_{h=1}^{n_i} \sigma^2(\boldsymbol{z}_{ih}' \boldsymbol{\gamma}),$$

which motivates us to estimate $\boldsymbol{\gamma}$ by solving the following estimating equation given by

$$\frac{1}{N} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left[\left\{ y_{ij} - \bar{y}_i - (\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)^t \widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} \right\}^2 - \left(1 - 2n_i^{-1}\right) \sigma^2(\boldsymbol{z}_{ij}^t \boldsymbol{\gamma}) - n_i^{-2} \sum_{h=1}^{n_i} \sigma^2(\boldsymbol{z}_{ih}^t \boldsymbol{\gamma}) \right] \boldsymbol{z}_{ij} = \boldsymbol{0},$$

which is equivalent to

$$\frac{1}{N} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left[\left\{ y_{ij} - \bar{y}_i - (\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)^t \widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} \right\}^2 \boldsymbol{z}_{ij} - \sigma^2(\boldsymbol{z}_{ij}^t \boldsymbol{\gamma})(\boldsymbol{z}_{ij} - 2n_i^{-1} \boldsymbol{z}_{ij} + n_i^{-1} \bar{\boldsymbol{z}}_i) \right] = \boldsymbol{0} \quad (5.9)$$

where $\bar{\boldsymbol{z}}_i = n_i^{-1} \sum_{j=1}^{n_i} \boldsymbol{z}_{ij}$. It is noted that, in the homoscedastic case with $\sigma^2(\boldsymbol{z}_{ij}^t \boldsymbol{\gamma}) = \delta^2$, the estimators of δ^2 and τ^2 reduce to the estimators identical to the Prasad-Rao estimator (Prasad and Rao, 1990) up to the constant factor.

Note that the function given in the left side of (5.9) does not depend on $\boldsymbol{\beta}$ and τ^2 and the estimator of τ^2 does not depend on $\boldsymbol{\beta}$ but on $\boldsymbol{\gamma}$. These suggest the simple algorithm for calculating the estimates of the model parameters: We first obtain the estimate $\widehat{\boldsymbol{\gamma}}$ of $\boldsymbol{\gamma}$ by solving (5.9), and then we get the estimate $\widehat{\tau}^2$ from (5.8) with $\boldsymbol{\gamma} = \widehat{\boldsymbol{\gamma}}$. Finally we have the GLS estimate $\widehat{\boldsymbol{\beta}}$ with substituting $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\tau}^2$ in (5.6).

5.2.3 Large sample properties

In this section, we provide large sample properties of the estimators given in the previous subsection when the number of clusters m goes to infinity, but n_i's are still bounded. To establish asymptotic results, we assume the following conditions under $m \to \infty$.

Assumption 5.1.

(A1) There exist bounded values $\underline{n}$ and $\overline{n}$ such that $\underline{n} \leq n_i \leq \overline{n}$ for $i = 1, \ldots, m$. The dimensions p and q are bounded, namely $p, q = O(1)$. The number of clusters with one observation, namely $n_i = 1$, is bounded.

(A2) The variance function $\sigma^2(\cdot)$ is twice differentiable and its derivatives are denoted by $(\sigma^2)^{(1)}(\cdot)$ and $(\sigma^2)^{(2)}(\cdot)$, respectively.

(A3) The following matrices converge to non-singular matrices:

$$m^{-1} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \boldsymbol{z}_{ij} \boldsymbol{z}_{ij}^t, \quad m^{-1} \sum_{i=1}^{m} \sum_{j=1}^{n_i} (\sigma^2)^{(a_1)}(\boldsymbol{z}_{ij}^t \boldsymbol{\gamma}) \boldsymbol{z}_{ij} \boldsymbol{z}_{ij}^t, \quad m^{-1} \boldsymbol{X}^t \boldsymbol{\Sigma}^{a_2} \boldsymbol{X}$$

for $a_1 = 1, 2$ and $a_2 = -1, 0, 1$.

(A4) $E[|v_i|^{8+c}] < \infty$ *and* $E[|\varepsilon_{ij}|^{8+c}] < \infty$ *for* $0 < c < 1$.

(A5) For all i *and* j, *there exist* $0 < \underline{c_1}, \overline{c_1} < \infty$ *and bounded values* $\underline{c_2}, \overline{c_2}$ *such that* $\underline{c_1} < \sigma^2(\boldsymbol{z}_{ij}^t\boldsymbol{\gamma}) < \overline{c_1}$ *and* $\underline{c_2} < (\sigma^2)^{(k)}(\boldsymbol{z}_{ij}^t\boldsymbol{\gamma}) < \overline{c_2}$ *with* $k = 1, 2$ *on the neighborhood of the true values.*

The conditions (A1) and (A3) are the standard assumptions in small area estimation. The condition (A2) is also non-restrictive, and the typical variance functions $\sigma^2(x) = x^2$ and $\sigma^2(x) = \exp x$ obviously satisfy the assumption. The moment condition (A4) is used for deriving second-order approximation of MSE of the EBLUP discussed in Section 5.3, and it is satisfied by many continuous distributions, including normal, shifted gamma, Laplace and t-distribution with degrees of freedom larger than 9. The three examples given in Section 5.2.1 satisfy the condition (A5).

In what follows, we use the notations

$$\sigma_{ij}^2 \equiv \sigma^2(\boldsymbol{z}_{ij}^t\boldsymbol{\gamma}), \quad \sigma_{ij(k)}^2 \equiv (\sigma^2)^{(k)}(\boldsymbol{z}_{ij}^t\boldsymbol{\gamma}), \quad k = 1, 2$$

for simplicity. To derive asymptotic approximations of the estimators, we use the following notations in the i-th cluster:

$$u_{1i} = \frac{m}{N}\sum_{j=1}^{n_i}\left\{(y_{ij} - \boldsymbol{x}_{ij}^t\boldsymbol{\beta})^2 - \sigma_{ij}^2 - \tau^2\right\}, \tag{5.10}$$

$$\boldsymbol{u}_{2i} = \frac{m}{N}\sum_{j=1}^{n_i}\left[\left\{y_{ij} - \bar{y}_i - (\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)^t\boldsymbol{\beta}\right\}^2\boldsymbol{z}_{ij} - \sigma_{ij}^2(\boldsymbol{z}_{ij} - 2n_i^{-1}\boldsymbol{z}_{ij} + n_i^{-1}\bar{\boldsymbol{z}}_i)\right], \tag{5.11}$$

with

$$\boldsymbol{T}_1(\boldsymbol{\gamma}) = \sum_{k=1}^{m}\sum_{h=1}^{n_k}\sigma_{kh(1)}^2\boldsymbol{z}_{kh}, \quad \boldsymbol{T}_2(\boldsymbol{\gamma}) = \left(\sum_{k=1}^{m}\sum_{j=1}^{n_k}\sigma_{kh(1)}^2(\boldsymbol{z}_{kh} - 2n_k^{-1}\boldsymbol{z}_{kh} + n_k^{-1}\bar{\boldsymbol{z}}_k)\boldsymbol{z}_{kh}'\right)^{-1}. \tag{5.12}$$

Note that $\boldsymbol{T}_1(\boldsymbol{\gamma}) = O(m)$ and $\boldsymbol{T}_2(\boldsymbol{\gamma}) = O(m^{-1})$ under Assumption 5.1. Then we obtain the asymptotically linear expression of the estimators.

Theorem 5.1. *Let* $\widehat{\boldsymbol{\phi}} = (\widehat{\boldsymbol{\beta}}', \widehat{\boldsymbol{\gamma}}', \widehat{\tau}^2)^t$ *be the estimator of* $\boldsymbol{\phi} = (\boldsymbol{\beta}^t, \boldsymbol{\gamma}^t, \tau^2)^t$. *Under Assumption 5.1, it holds that* $\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi} = O_p(m^{-1/2})$ *with the asymptotically linear expression*

$$\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi} = \frac{1}{m}\sum_{i=1}^{m}((\boldsymbol{\psi}_i^{\boldsymbol{\beta}})^t, (\boldsymbol{\psi}_i^{\boldsymbol{\gamma}})^t, \psi_i^{\tau})^t + o_p(m^{-1/2}),$$

where

$$\boldsymbol{\psi}_i^{\boldsymbol{\beta}} = m\left(\boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}_i\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}), \quad \boldsymbol{\psi}_i^{\boldsymbol{\gamma}} = N\boldsymbol{T}_2(\boldsymbol{\gamma})\boldsymbol{u}_{2i}, \quad \psi_i^{\tau} = u_{1i} - \boldsymbol{T}_1(\boldsymbol{\gamma})^t\boldsymbol{T}_2(\boldsymbol{\gamma})\boldsymbol{u}_{2i}.$$

From Theorem 5.1, it follows that $m^{1/2}(\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi})$ has an asymptotically normal distribution with mean vector $\mathbf{0}$ and covariance matrix $m\boldsymbol{\Omega}$, where $\boldsymbol{\Omega}$ is a $(p+q+1)\times(p+q+1)$ matrix partitioned as

$$m\boldsymbol{\Omega} \equiv \begin{pmatrix} m\boldsymbol{\Omega}_{\boldsymbol{\beta}\boldsymbol{\beta}} & m\boldsymbol{\Omega}_{\boldsymbol{\beta}\boldsymbol{\gamma}} & m\boldsymbol{\Omega}_{\boldsymbol{\beta}\tau} \\ m\boldsymbol{\Omega}'_{\boldsymbol{\beta}\boldsymbol{\gamma}} & m\boldsymbol{\Omega}_{\boldsymbol{\gamma}\boldsymbol{\gamma}} & m\boldsymbol{\Omega}_{\boldsymbol{\gamma}\tau} \\ m\boldsymbol{\Omega}'_{\boldsymbol{\beta}\tau} & m\boldsymbol{\Omega}'_{\boldsymbol{\gamma}\tau} & m\Omega_{\tau\tau} \end{pmatrix} = \lim_{m\to\infty}\frac{1}{m}\sum_{i=1}^{m}\begin{pmatrix} E[\psi_i^{\boldsymbol{\beta}}\psi_i^{\boldsymbol{\beta}^t}] & E[\psi_i^{\boldsymbol{\beta}}\psi_i^{\boldsymbol{\gamma}^t}] & E[\psi_i^{\boldsymbol{\beta}}\psi_i^{\tau}] \\ E[\psi_i^{\boldsymbol{\gamma}}\psi_i^{\boldsymbol{\beta}^t}] & E[\psi_i^{\boldsymbol{\gamma}}\psi_i^{\boldsymbol{\gamma}^t}] & E[\psi_i^{\boldsymbol{\gamma}}\psi_i^{\tau}] \\ E[\psi_i^{\tau}\psi_i^{\boldsymbol{\beta}^t}] & E[\psi_i^{\tau}\psi_i^{\boldsymbol{\gamma}^t}] & E[\psi_i^{\tau}\psi_i^{\tau}] \end{pmatrix}.$$

It is noticed that $E[u_{1i}(y_{ij}-\boldsymbol{x}_{ij}^t\boldsymbol{\beta})]=0$ and $E[u_{2i}(y_{ij}-\boldsymbol{x}_{ij}^t\boldsymbol{\beta})]=0$ when y_{ij} are normally distributed. In such a case, it follows $\boldsymbol{\Omega}_{\boldsymbol{\beta}\boldsymbol{\gamma}}=0$ and $\boldsymbol{\Omega}_{\boldsymbol{\beta}\tau}=\mathbf{0}$, namely $\boldsymbol{\beta}$ and $\boldsymbol{\phi}=(\boldsymbol{\gamma}^t,\tau^2)^t$ are asymptotically orthogonal. However, since we do not assume the normality for observations y_{ij}'s, $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$ are not necessarily orthogonal.

The asymptotic covariance matrix $m\boldsymbol{\Omega}$ or $\boldsymbol{\Omega}$ can be easily estimated from samples. For example, $m\boldsymbol{\Omega}_{\boldsymbol{\beta}\boldsymbol{\beta}}=\lim_{m\to\infty} m^{-1}\sum_{i=1}^{m}E[\psi_i^{\boldsymbol{\beta}}\psi_i^{\boldsymbol{\beta}^t}]$ can be estimated by

$$m\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\beta}\boldsymbol{\beta}}=\frac{1}{m}\sum_{i=1}^{m}\widehat{\psi_i^{\boldsymbol{\beta}}}\widehat{\psi_i^{\boldsymbol{\beta}^t}},$$

where $\widehat{\psi_i^{\boldsymbol{\beta}}}$ is obtained by replacing unknown parameters $\boldsymbol{\phi}$ in $\psi_i^{\boldsymbol{\beta}}$ with estimates $\widehat{\boldsymbol{\phi}}$. It is noted that the accuracy of estimation is given by

$$\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\beta}\boldsymbol{\beta}}=\boldsymbol{\Omega}_{\boldsymbol{\beta}\boldsymbol{\beta}}+o_p(m^{-1}),$$

from Theorem 5.1 and $\boldsymbol{\Omega}=O(m^{-1})$. The estimator $\widehat{\boldsymbol{\Omega}}$ will be used to get the estimators of mean squared errors of predictors in Section 5.3.

We next provide the asymptotic properties of conditional covariance matrix given in the following corollary where the proof is given in Section 5.5.

Corollary 5.1. *Under Assumption 5.1, for $i=1,\ldots,m$, it follows that*

$$E\left((\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi})(\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi})^t\Big|\boldsymbol{y}_i\right)=\boldsymbol{\Omega}+c(\boldsymbol{y}_i)o(m^{-1}), \tag{5.13}$$

where $c(\boldsymbol{y}_i)$ is the fourth-order function of $\boldsymbol{y}_i$, so that $E|c(\boldsymbol{y}_i)|<\infty$ under Assumption 5.1.

This property is used for estimation and evaluating the mean squared errors of EBLUP discussed in the subsequent section. Moreover, in the evaluation of the mean squared errors of EBLUP and the derivation of its estimators, we need to obtain the conditional and unconditional asymptotic biases of the estimators $\widehat{\boldsymbol{\phi}}$.

Let $\boldsymbol{b}_{\boldsymbol{\beta}}^{(i)}(\boldsymbol{y}_i),\boldsymbol{b}_{\boldsymbol{\gamma}}^{(i)}(\boldsymbol{y}_i)$ and $\boldsymbol{b}_{\tau}^{(i)}(\boldsymbol{y}_i)$ be the second-order conditional asymptotic biases defined as

$$E[\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}|\boldsymbol{y}_i]=\boldsymbol{b}_{\boldsymbol{\beta}}^{(i)}(\boldsymbol{y}_i)+o_p(m^{-1}),\quad E[\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}|\boldsymbol{y}_i]=\boldsymbol{b}_{\boldsymbol{\gamma}}^{(i)}(\boldsymbol{y}_i)+o_p(m^{-1}),$$
$$E[\widehat{\tau}^2-\tau^2|\boldsymbol{y}_i]=\boldsymbol{b}_{\tau}^{(i)}(\boldsymbol{y}_i)+o_p(m^{-1}).$$

In the following theorem, we provide the analytical expressions of $\boldsymbol{b}_{\boldsymbol{\beta}}^{(i)}(\boldsymbol{y}_i), \boldsymbol{b}_{\boldsymbol{\gamma}}^{(i)}(\boldsymbol{y}_i)$ and $\boldsymbol{b}_{\tau}^{(i)}(\boldsymbol{y}_i)$. Define $\boldsymbol{b}_{\boldsymbol{\beta}}$, $\boldsymbol{b}_{\boldsymbol{\gamma}}$ and b_τ by

$$\begin{aligned}\boldsymbol{b}_{\boldsymbol{\beta}} = \left(\boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right)^{-1}\Bigg\{&\sum_{s=1}^{q}\sum_{k=1}^{m}\boldsymbol{X}_k^t\boldsymbol{\Sigma}_k^{-1}\boldsymbol{W}_{i(s)}\boldsymbol{\Sigma}_k^{-1}\boldsymbol{X}_k\left(\boldsymbol{\Omega}_{\beta^*\gamma_s}-\boldsymbol{\Omega}_{\beta\gamma_s}\right)\\&+\sum_{k=1}^{m}\boldsymbol{X}_k^t\boldsymbol{\Sigma}_k^{-1}\boldsymbol{J}_{n_k}\boldsymbol{\Sigma}_k^{-1}\boldsymbol{X}_k(\boldsymbol{\Omega}_{\boldsymbol{\beta}^*\tau}-\boldsymbol{\Omega}_{\boldsymbol{\beta}\tau})\Bigg\}\end{aligned}$$

$$\begin{aligned}\boldsymbol{b}_{\boldsymbol{\gamma}} = \boldsymbol{T}_2(\boldsymbol{\gamma})\Bigg[&2\sum_{k=1}^{m}\mathrm{col}\left\{\mathrm{tr}\left(\mathrm{E}_k\boldsymbol{Z}_{kr}\mathrm{E}_k\boldsymbol{X}_k\left[\boldsymbol{V}_{\mathrm{OLS}}\boldsymbol{X}_k^t-(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}_k^t\boldsymbol{\Sigma}_k\right]\right)\right\}_r\\&-\sum_{k=1}^{m}\sum_{j=1}^{n_k}\boldsymbol{z}_{kj}\sigma_{kj(2)}^2(\boldsymbol{z}_{kj}-2n_k^{-1}\boldsymbol{z}_{kj}+n_k^{-1}\bar{\boldsymbol{z}}_k)^t\boldsymbol{\Omega}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}\boldsymbol{z}_{kj}\Bigg],\end{aligned}\tag{5.14}$$

and

$$\begin{aligned}b_\tau = &-\frac{1}{N}\sum_{k=1}^{m}\sum_{j=1}^{n_k}\sigma_{kj(1)}^2\boldsymbol{z}_{kj}^t\boldsymbol{b}_{\boldsymbol{\gamma}}-\frac{2}{N}\sum_{k=1}^{m}\mathrm{tr}\left\{(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}_k^t\boldsymbol{\Sigma}_k\boldsymbol{X}_k\right\}\\&-\frac{1}{2N}\sum_{k=1}^{m}\sum_{j=1}^{n_k}\sigma_{kj(2)}^2\boldsymbol{z}_{kj}^t\boldsymbol{\Omega}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}\boldsymbol{z}_{kj}+\frac{1}{N}\sum_{k=1}^{m}\mathrm{tr}\left(\boldsymbol{X}_k^t\boldsymbol{X}_k\boldsymbol{V}_{\mathrm{OLS}}\right),\end{aligned}$$

where $\mathrm{E}_k = \boldsymbol{I}_{n_k} - n_k^{-1}\boldsymbol{J}_{n_k}$, $\boldsymbol{V}_{\mathrm{OLS}} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{\Sigma}\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}$, $\boldsymbol{Z}_{kr} = \mathrm{diag}(z_{k1r},\ldots,z_{kn_kr})$ for r-th element z_{kjr} of $\boldsymbol{z}_{kj}$, $\boldsymbol{\Omega}_{\boldsymbol{\beta}^*a}$ for $a \in \{\tau,\gamma_1,\ldots,\gamma_q\}$ and $\boldsymbol{W}_{i(s)}$ are defined in the proof of Theorem 5.2, and $\mathrm{col}\{a_r\}_r$ denotes a q-dimensional vector $(a_1,\ldots,a_q)^t$. It is noted that $\boldsymbol{b}_{\boldsymbol{\beta}}, \boldsymbol{b}_{\boldsymbol{\gamma}}, b_\tau$ are of order $O(m^{-1})$. Now we provide the second-order approximation to the conditional asymptotic bias.

Theorem 5.2. *Under Assumption 5.1, we have*

$$\begin{aligned}&\boldsymbol{b}_{\boldsymbol{\beta}}^{(i)}(\boldsymbol{y}_i) = \left(\boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}_i^t\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta})+\boldsymbol{b}_{\boldsymbol{\beta}}, \quad \boldsymbol{b}_{\boldsymbol{\gamma}}^{(i)}(\boldsymbol{y}_i) = \boldsymbol{T}_2(\boldsymbol{\gamma})\boldsymbol{u}_{2i}+\boldsymbol{b}_{\boldsymbol{\gamma}}\\&b_\tau^{(i)}(\boldsymbol{y}_i) = m^{-1}u_{1i}-m^{-1}\boldsymbol{T}_1(\boldsymbol{\gamma})^t\boldsymbol{T}_2(\boldsymbol{\gamma})\boldsymbol{u}_{2i}+b_\tau,\end{aligned}\tag{5.15}$$

where $\boldsymbol{b}_{\boldsymbol{\beta}}^{(i)}(\boldsymbol{y}_i)$, $\boldsymbol{b}_{\boldsymbol{\gamma}}^{(i)}(\boldsymbol{y}_i)$ and $b_\tau^{(i)}(\boldsymbol{y}_i)$ are of order $O_p(m^{-1})$, and u_{1i} and u_{2i} are given in (5.10) *and* (5.11)*, respectively.*

From the above theorem, we immediately obtain the unconditional asymptotic bias of the estimators $\widehat{\boldsymbol{\phi}}$ by taking expectation with respect to $\boldsymbol{y}_i$ given in the following Corollary.

Corollary 5.2. *Under Assumption 5.1, it holds that*

$$E[\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi}] = (\boldsymbol{b}_{\boldsymbol{\beta}}^t, \boldsymbol{b}_{\boldsymbol{\gamma}}^t, b_\tau)^t + o(m^{-1}),$$

where $\boldsymbol{b}_{\boldsymbol{\beta}}$, $\boldsymbol{b}_{\boldsymbol{\gamma}}$ and b_τ are given in (5.14).

5.3 Prediction and Risk Evaluation

5.3.1 Empirical predictor

We now consider the prediction of

$$\mu_i = \boldsymbol{c}_i^t\boldsymbol{\beta} + v_i,$$

where $\boldsymbol{c}_i$ is a known (user specified) vector and v_i is the random effect in model (5.3). The typical choice of $\boldsymbol{c}_i$ is $\boldsymbol{c}_i = \bar{\boldsymbol{x}}_i$ which corresponds to the prediction of mean of the i-th cluster. A predictor $\widetilde{\mu}(\boldsymbol{y}_i)$ of μ_i is evaluated in terms of the MSE $E[(\widetilde{\mu}(\boldsymbol{y}_i)-\mu_i)^2]$. In the general forms of $\widetilde{\mu}(\boldsymbol{y}_i)$, the minimizer (best predictor) of the MSE cannot be obtain without a distributional assumption for v_i and ε_{ij}. Thus we focus on the class of linear and unbiased predictors, and the best linear unbiased predictor (BLUP) of μ_i in terms of the MSE is given by

$$\widetilde{\mu}_i = \boldsymbol{c}_i^t\boldsymbol{\beta} + \mathbf{1}_{n_i}^t\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}).$$

This can be simplified as

$$\widetilde{\mu}_i = \boldsymbol{c}_i^t\boldsymbol{\beta} + \sum_{j=1}^{n_i}\lambda_{ij}\left(y_{ij} - \boldsymbol{x}_{ij}^t\boldsymbol{\beta}\right),$$

where $\lambda_{ij} = \tau^2\sigma_{ij}^{-2}\eta_i^{-1}$ for $\eta_i = 1 + \tau^2\sum_{h=1}^{n_i}\sigma_{ih}^{-2}$. In the case of homogeneous variances, namely $\sigma_{ij}^2 = \delta^2$, it is confirmed that the BLP reduces to $\widetilde{\mu}_i = \boldsymbol{c}_i^t\boldsymbol{\beta} + \lambda_i\left(\bar{y}_i - \bar{\boldsymbol{x}}_i^t\boldsymbol{\beta}\right)$ with $\lambda_i = n_i\tau^2(\delta^2 + n_i\tau^2)^{-1}$. The BLUP is not feasible since it depends on unknown parameters $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and τ^2. Plugging the estimators into $\widetilde{\mu}_i$, we get the empirical best linear unbiased predictor (EBLUP)

$$\widehat{\mu}_i = \boldsymbol{c}_i^t\widehat{\boldsymbol{\beta}} + \sum_{j=1}^{n_i}\widehat{\lambda}_{ij}\left(y_{ij} - \boldsymbol{x}_{ij}^t\widehat{\boldsymbol{\beta}}\right), \qquad \widehat{\lambda}_{ij} = \widehat{\tau}^2\widehat{\sigma}_{ij}^{-2}\widehat{\eta}_i^{-1} \tag{5.16}$$

for $\widehat{\eta}_i^{-1} = 1 + \widehat{\tau}^2\sum_{h=1}^{n_i}\widehat{\sigma}_{ih}^{-2}$. In the subsequent section, we consider the mean squared errors (MSE) of EBLUP (5.16) without any distributional assumptions for v_i and ε_{ij}.

5.3.2 Second-order approximation to MSE

To evaluate uncertainty of EBLUP given by (5.16), we evaluate the MSE defined as $\mathrm{MSE}_i(\boldsymbol{\phi}) = E\left[(\widehat{\mu}_i - \mu_i)^2\right]$ for $\boldsymbol{\phi} = (\boldsymbol{\gamma}^t, \tau^2)^t$. The MSE is decomposed as

$$\begin{aligned}\mathrm{MSE}_i(\boldsymbol{\phi}) &= E\left[(\widehat{\mu}_i - \widetilde{\mu}_i + \widetilde{\mu}_i - \mu_i)^2\right] \\ &= E\left[(\widetilde{\mu}_i - \mu_i)^2\right] + E\left[(\widehat{\mu}_i - \widetilde{\mu}_i)^2\right] + 2E\left[(\widehat{\mu}_i - \widetilde{\mu}_i)(\widetilde{\mu}_i - \mu_i)\right].\end{aligned}$$

From the expression of $\widetilde{\mu}_i$, we have

$$\widetilde{\mu}_i - \mu_i = \left(\sum_{j=1}^{n_i}\lambda_{ij} - 1\right)v_i + \sum_{j=1}^{n_i}\lambda_{ij}\varepsilon_{ij},$$

which leads to

$$R_{1i}(\boldsymbol{\phi}) \equiv E\left[(\widetilde{\mu}_i - \mu_i)^2\right] = \left(\sum_{j=1}^{n_i}\lambda_{ij} - 1\right)^2\tau^2 + \sum_{j=1}^{n_i}\lambda_{ij}^2\sigma_{ij}^2 = \tau^2\eta_i^{-1}. \tag{5.17}$$

For the second term, however, we cannot obtain an exact expression, so that we derive the approximation up to $O(m^{-1})$. Using the Taylor series expansion, we have

$$\widehat{\mu}_i - \widetilde{\mu}_i = \left(\frac{\partial \widetilde{\mu}_i}{\partial \boldsymbol{\phi}}\right)^t (\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}) + \frac{1}{2}(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi})^t \left(\frac{\partial^2 \widetilde{\mu}_i}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}^t}\Big|_{\boldsymbol{\phi}=\boldsymbol{\phi}^*}\right)(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}), \tag{5.18}$$

where $\boldsymbol{\phi}^*$ is on the line between $\boldsymbol{\phi}$ and $\widehat{\boldsymbol{\phi}}$. The straightforward calculation shows that

$$\frac{\partial \widetilde{\mu}_i}{\partial \boldsymbol{\beta}} = \boldsymbol{c}_i - \sum_{j=1}^{n_i} \lambda_{ij} \boldsymbol{x}_{ij}, \quad \frac{\partial \widetilde{\mu}_i}{\partial \boldsymbol{\gamma}} = \eta_i^{-2} \sum_{j=1}^{n_i} \sigma_{ij}^{-2} \boldsymbol{\delta}_{ij} (y_{ij} - \boldsymbol{x}_{ij}^t \boldsymbol{\beta}), \quad \frac{\partial \widetilde{\mu}_i}{\partial \tau^2} = \eta_i^{-2} \sum_{j=1}^{n_i} \sigma_{ij}^{-2} (y_{ij} - \boldsymbol{x}_{ij}^t \boldsymbol{\beta}), \tag{5.19}$$

where

$$\boldsymbol{\delta}_{ij} = \tau^4 \sum_{h=1}^{n_i} \sigma_{ih}^{-4} \sigma_{ih(1)}^2 \boldsymbol{z}_{ih} - \tau^2 \eta_i \sigma_{ij}^{-2} \sigma_{ij(1)}^2 \boldsymbol{z}_{ij}.$$

Then each element in $\partial^2 \widetilde{\mu}_i / \partial \boldsymbol{\phi} \partial \boldsymbol{\phi}^t$ is a linear function of $\boldsymbol{y}_i$. Hence under Assumption 5.1, using the similar arguments given in Lahiri and Rao (1995), we can show that

$$E\left[(\widehat{\mu}_i - \widetilde{\mu}_i)^2\right] = R_{2i}(\boldsymbol{\phi}) + o(m^{-1}), \tag{5.20}$$

where the detailed proof is given in Section 5.5, and

$$\begin{aligned} R_{2i}(\boldsymbol{\phi}) =& \eta_i^{-4} \tau^2 \left(\sum_{j=1}^{n_i} \sigma_{ij}^{-2} \boldsymbol{\delta}_{ij}\right)^t \boldsymbol{\Omega}_{\gamma\gamma} \left(\sum_{j=1}^{n_i} \sigma_{ij}^{-2} \boldsymbol{\delta}_{ij}\right) + \eta_i^{-4} \sum_{j=1}^{n_i} \sigma_{ij}^{-2} \boldsymbol{\delta}_{ij}^t \boldsymbol{\Omega}_{\gamma\gamma} \boldsymbol{\delta}_{ij} \\ &+ 2\eta_i^{-3} \sum_{j=1}^{n_i} \sigma_{ij}^{-2} \boldsymbol{\delta}_{ij}^t \boldsymbol{\Omega}_{\gamma\tau} + \eta_i^{-3} \sum_{j=1}^{n_i} \sigma_{ij}^{-2} \Omega_{\tau\tau} + \left(\boldsymbol{c}_i - \sum_{j=1}^{n_i} \lambda_{ij} \boldsymbol{x}_{ij}\right)^t \boldsymbol{\Omega}_{\boldsymbol{\beta}\boldsymbol{\beta}} \left(\boldsymbol{c}_i - \sum_{j=1}^{n_i} \lambda_{ij} \boldsymbol{x}_{ij}\right), \end{aligned} \tag{5.21}$$

which is of order $O(m^{-1})$. All the evaluations of the residual terms appeared in this paper can be done by the similar manner, and detailed proofs will be omitted in what follows.

We next evaluate the cross term $E\left[(\widehat{\mu}_i - \widetilde{\mu}_i)(\widetilde{\mu}_i - \mu_i)\right]$. This term vanishes under the normality assumptions for v_i and ε_{ij}, but in general, it cannot be neglected. As in the case of R_{2i}, we obtain an approximation of $E\left[(\widehat{\mu}_i - \widetilde{\mu}_i)(\widetilde{\mu}_i - \mu_i)\right]$ up to $O(m^{-1})$. Noting that

$$\widetilde{\mu}_i - \mu_i = \left(\sum_{j=1}^{n_i} \lambda_{ij} - 1\right) v_i + \sum_{j=1}^{n_i} \lambda_{ij} \varepsilon_{ij} \equiv w_i,$$

and using the expansion (5.18), we obtain

$$E\left[(\widehat{\mu}_i - \widetilde{\mu}_i)(\widetilde{\mu}_i - \mu_i)\right] = E\left[\left(\frac{\partial \widetilde{\mu}_i}{\partial \boldsymbol{\phi}}\right)^t (\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}) w_i\right] + \frac{1}{2} E\left[(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi})^t \left(\frac{\partial^2 \widetilde{\mu}_i}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}^t}\Big|_{\boldsymbol{\phi}=\boldsymbol{\phi}^*}\right)(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}) w_i\right].$$

Using the expression of (5.19) and Corollary 5.1, the straightforward calculation (whose details are given in Section 5.5) shows that

$$R_{32i}(\boldsymbol{\phi}) \equiv \frac{1}{2} E\left[(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi})^t \left(\frac{\partial^2 \widetilde{\mu}_i}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}^t}\Big|_{\boldsymbol{\phi}=\boldsymbol{\phi}^*}\right)(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}) w_i\right] = o(m^{-1}),$$

under Assumption 5.1. Moreover, from Theorem 5.2, we obtain

$$E\left[\left(\frac{\partial\widetilde{\mu}_i}{\partial\boldsymbol{\phi}}\right)^t(\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi})w_i\right] = R_{31i}(\boldsymbol{\phi},\boldsymbol{\kappa}) + o(m^{-1}),$$

for

$$\begin{aligned} R_{31i}(\boldsymbol{\phi},\boldsymbol{\kappa}) = \eta_i^{-2}\sum_{j=1}^{n_i}\sigma_{ij}^{-2}\boldsymbol{\delta}_{ij}^t\left(\sum_{k=1}^{m}\sum_{h=1}^{n_k}\sigma_{kh(1)}^2\boldsymbol{z}_{kh}\boldsymbol{z}_{kh}^t\right)^{-1}\boldsymbol{M}_{2ij}(\boldsymbol{\phi},\boldsymbol{\kappa}) \\ + m^{-1}\eta_i^{-2}\sum_{j=1}^{n_i}\sigma_{ij}^{-2}\Big\{M_{1ij}(\boldsymbol{\phi},\boldsymbol{\kappa}) - \boldsymbol{T}_1(\boldsymbol{\gamma})^t\boldsymbol{T}_2(\boldsymbol{\gamma})\boldsymbol{M}_{2ij}(\boldsymbol{\phi},\boldsymbol{\kappa})\Big\}, \end{aligned} \tag{5.22}$$

where

$$\begin{aligned} M_{1ij}(\boldsymbol{\phi},\boldsymbol{\kappa}) &= mN^{-1}\tau^2\eta_i^{-1}\Big\{n_i\tau^2(3-\kappa_v) + \sigma_{ij}^2(\kappa_\varepsilon - 3)\Big\} \\ \boldsymbol{M}_{2ij}(\boldsymbol{\phi},\boldsymbol{\kappa}) &= mN^{-1}\tau^2\eta_i^{-1}n_i^{-2}(n_i-1)^2(\kappa_\varepsilon-3)\sigma_{ij}^2\boldsymbol{z}_{ij}, \end{aligned}$$

and κ_v, κ_ε are defined as $E(v_i^4) = \kappa_v\tau^4$ and $E(\varepsilon_{ij}^4) = \kappa_\varepsilon\sigma_{ij}^4$, respectively, and $\boldsymbol{\kappa} = (\kappa_v, \kappa_\varepsilon)^t$. The derivation of the expression of $R_{31i}(\boldsymbol{\phi},\boldsymbol{\kappa})$ is also given in Section 5.5. From the expression (5.22), it holds that $R_{31i}(\boldsymbol{\phi},\boldsymbol{\kappa}) = O(m^{-1})$.

Under the normality assumption of v_i and ε_{ij}, we immediately obtain $M_{1ij} = 0$ and $\boldsymbol{M}_{2ij} = \boldsymbol{0}$ since $\boldsymbol{\kappa} = (3,3)^t$. This leads to $R_{31} = 0$, which means that the cross term does not appear in the second-order approximated MSE, that is our result is consistent to the well-known result.

Now, we summarize the result for the second-order approximation of the MSE.

Theorem 5.3. *Under Assumption 5.1, the second-order approximation of the MSE is given by*

$$\mathrm{MSE}_i(\boldsymbol{\phi}) = R_{1i}(\boldsymbol{\phi}) + R_{2i}(\boldsymbol{\phi}) + 2R_{31i}(\boldsymbol{\phi},\boldsymbol{\kappa}) + o(m^{-1}),$$

where $R_{1i}(\boldsymbol{\phi})$, $R_{2i}(\boldsymbol{\phi})$ and $R_{31i}(\boldsymbol{\phi},\boldsymbol{\kappa})$ are given in (5.17), (5.21) *and* (5.22)*, respectively, and $R_{1i}(\boldsymbol{\phi}) = O(1)$, $R_{2i}(\boldsymbol{\phi}) = O(m^{-1})$ and $R_{31i}(\boldsymbol{\phi},\boldsymbol{\kappa}) = O(m^{-1})$.*

The approximated MSE given in Theorem 5.3 depends on unknown parameters. Thus, in the subsequent section, we derive the second-order unbiased estimator of the MSE by the analytical and the matching bootstrap methods.

5.3.3 Analytical estimator of the MSE

We first derive the analytical second-order unbiased estimator of the MSE. From Theorem 5.3, $R_{2i}(\boldsymbol{\phi})$ is $O(m^{-1})$, so that it can be estimated by the plug-in estimator $R_{2i}(\widehat{\boldsymbol{\phi}})$ with second-order accuracy, namely $E[R_{2i}(\widehat{\boldsymbol{\phi}})] = R_{2i}(\boldsymbol{\phi}) + o(m^{-1})$. For $R_{31i}(\boldsymbol{\phi},\boldsymbol{\kappa})$ with order $O(m^{-1})$, if a consistent estimator $\widehat{\boldsymbol{\kappa}}$ is available for $\boldsymbol{\kappa}$, this term can be estimated by the plug-in estimator with second-order unbiasedness. To this end, we construct a consistent estimator of $\boldsymbol{\kappa}$ using

the expression of fourth moment of observations. The straightforward calculation shows that

$$E\left[\sum_{j=1}^{n_i}\left\{y_{ij}-\bar{y}_i-(\boldsymbol{x}_{ij}-\bar{\boldsymbol{x}}_i)^t\boldsymbol{\beta}\right\}^4\right]$$
$$=\kappa_\varepsilon n_i^{-4}(n_i-1)(n_i-2)(n_i^2-n_i-1)\left(\sum_{j=1}^{n_i}\sigma_{ij}^4\right)+3n_i^{-3}(2n_i-3)\left\{\left(\sum_{j=1}^{n_i}\sigma_{ij}^2\right)^2-\sum_{j=1}^{n_i}\sigma_{ij}^4\right\},$$

whereby we can estimate κ_ε by

$$\widehat{\kappa}_\varepsilon=\frac{1}{N^*}\sum_{i=1}^{m}\left[\sum_{j=1}^{n_i}\left\{y_{ij}-\bar{y}_i-(\boldsymbol{x}_{ij}-\bar{\boldsymbol{x}}_i)^t\widehat{\boldsymbol{\beta}}\right\}^4-3n_i^{-3}(2n_i-3)\left\{\left(\sum_{j=1}^{n_i}\sigma_{ij}^2\right)^2-\sum_{j=1}^{n_i}\sigma_{ij}^4\right\}\right], \tag{5.23}$$

where $N^*=n_i^{-4}(n_i-1)(n_i-2)(n_i^2-n_i-1)\sum_{j=1}^{n_i}\sigma_{ij}^4$ and $\widehat{\boldsymbol{\beta}}$ is the feasible GLS estimator of $\boldsymbol{\beta}$. For κ_v, it is observed that

$$E\left[\left(y_{ij}-\boldsymbol{x}_{ij}^t\boldsymbol{\beta}\right)^4\right]=\tau^4\kappa_v+6\tau^2\sigma_{ij}^2+\kappa_\varepsilon\sigma_{ij}^4,$$

which leads to the estimator of κ_v given by

$$\widehat{\kappa}_v=\frac{1}{N\widehat{\tau}^4}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left\{\left(y_{ij}-\boldsymbol{x}_{ij}^t\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}\right)^4-6\widehat{\tau}^2\widehat{\sigma}_{ij}^2-\widehat{\kappa}_\varepsilon\widehat{\sigma}_{ij}^4\right\}. \tag{5.24}$$

From Theorem 5.1, it immediately follows that the estimators given in (5.23) and (5.24) are consistent. Using these estimators, we can estimate R_{31i} by $R_{31i}(\widehat{\boldsymbol{\phi}},\widehat{\boldsymbol{\kappa}})$ with second-order accuracy.

Finally, we consider the second-order unbiased estimation of R_{1i}. The situation is different than before since $R_{1i}=O(1)$, which means that the plug-in estimator $R_{1i}(\widehat{\boldsymbol{\phi}})$ has the second-order bias with $O(m^{-1})$. Thus we need to obtain the second-order bias of $R_{1i}(\widehat{\boldsymbol{\phi}})$ and correct them. By the Taylor series expansion, we have

$$R_{1i}(\widehat{\boldsymbol{\phi}})=R_{1i}(\boldsymbol{\phi})+\left(\frac{\partial R_{1i}(\boldsymbol{\phi})}{\partial\boldsymbol{\phi}^t}\right)(\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi})+\frac{1}{2}(\boldsymbol{\phi}-\boldsymbol{\phi})^t\left(\frac{\partial^2 R_{1i}(\boldsymbol{\phi})}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}^t}\right)(\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi})+o_p(\|\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi}\|^2).$$

Then, the second-order bias of $R_{1i}(\widehat{\boldsymbol{\phi}})$ is expressed as

$$\begin{aligned}E[R_{1i}(\widehat{\boldsymbol{\phi}})]-R_{1i}(\boldsymbol{\phi})&\\&=\left(\frac{\partial R_{1i}(\boldsymbol{\phi})}{\partial\boldsymbol{\phi}^t}\right)E[\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi}]+\frac{1}{2}\mathrm{tr}\left\{\left(\frac{\partial^2 R_{1i}(\boldsymbol{\phi})}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}^t}\right)E\left[(\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi})(\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi})^t\right]\right\}+o(m^{-1})\\&=\left(\frac{\partial R_{1i}(\boldsymbol{\phi})}{\partial\boldsymbol{\phi}^t}\right)\boldsymbol{b}_{\boldsymbol{\phi}}+\frac{1}{2}\mathrm{tr}\left\{\left(\frac{\partial^2 R_{1i}(\boldsymbol{\phi})}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}^t}\right)\boldsymbol{\Omega}_{\boldsymbol{\phi}}\right\}+o(m^{-1}),\end{aligned}$$

where $\boldsymbol{\Omega}_{\boldsymbol{\phi}}$ is the sub-matrix of $\boldsymbol{\Omega}$ with respect to $\boldsymbol{\phi}$, and $\boldsymbol{b}_{\phi}$ is the second-order bias of $\widehat{\boldsymbol{\phi}}$ given in Corollary 5.2. The straightforward calculation shows that

$$\frac{\partial R_{1i}(\boldsymbol{\phi})}{\partial \tau^2} = \eta_i^{-2}, \quad \frac{\partial R_{1i}(\boldsymbol{\phi})}{\partial \boldsymbol{\gamma}} = -\tau^2\eta_i^{-2}\boldsymbol{\eta}_{i(1)}, \quad \frac{\partial^2 R_{1i}(\boldsymbol{\phi})}{\partial \tau^2 \partial \tau^2} = 2\tau^{-2}(\eta_i^{-3} - \eta_i^{-2}),$$
$$\frac{\partial^2 R_{1i}(\boldsymbol{\phi})}{\partial \boldsymbol{\gamma} \partial \tau^2} = -2\eta_i^{-3}\boldsymbol{\eta}_{i(1)}, \quad \frac{\partial^2 R_{1i}(\boldsymbol{\phi})}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^t} = \tau^2\eta_i^{-3}(2\boldsymbol{\eta}_{i(1)}\boldsymbol{\eta}_{i(1)}^t - \eta_i\boldsymbol{\eta}_{i(2)}),$$

where

$$\boldsymbol{\eta}_{i(1)} \equiv \frac{\partial \eta_i}{\partial \boldsymbol{\gamma}} = -\tau^2 \sum_{j=1}^{n_i} \sigma_{ij}^{-4}\sigma_{ij(1)}^2 \boldsymbol{z}_{ij}, \quad \boldsymbol{\eta}_{i(2)} \equiv \frac{\partial^2 \eta_i}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^t} = \tau^2 \sum_{j=1}^{n_i} \left(2\sigma_{ij}^{-2}\sigma_{ij(1)}^4 - \sigma_{ij(2)}^2\right) \sigma_{ij}^{-4}\boldsymbol{z}_{ij}\boldsymbol{z}_{ij}^t.$$

Therefore, we obtain the expression of the second-order bias given by

$$\begin{aligned} B_i(\boldsymbol{\phi}) = &- \tau^2\eta_i^{-2}\boldsymbol{\eta}_{i(1)}^t\boldsymbol{b}_{\gamma} + \eta_i^{-2}b_{\tau} - 2\eta_i^{-3}\boldsymbol{\eta}_{i(1)}^t\boldsymbol{\Omega}_{\gamma\tau} + \tau^{-2}(\eta_i^{-3} - \eta_i^{-2})\Omega_{\tau\tau} \\ &+ \tau^2\eta_i^{-3}\left\{\boldsymbol{\eta}_{i(1)}^t\boldsymbol{\Omega}_{\gamma\gamma}\boldsymbol{\eta}_{i(1)} - \frac{1}{2}\eta_i \mathrm{tr}\left(\boldsymbol{\eta}_{i(2)}\boldsymbol{\Omega}_{\gamma\gamma}\right)\right\}, \end{aligned} \tag{5.25}$$

with $B_i(\boldsymbol{\phi}) = O(m^{-1})$. Noting that $B_i(\boldsymbol{\phi})$ can be estimated by $B_i(\widehat{\boldsymbol{\phi}})$ with $E[B_i(\widehat{\boldsymbol{\phi}})] = B_i(\boldsymbol{\phi}) + o(m^{-1})$ from Theorem 5.1, we propose the bias corrected estimator of R_{1i} given by

$$\widehat{R_{1i}}(\widehat{\boldsymbol{\phi}})^{bc} = R_{1i}(\widehat{\boldsymbol{\phi}}) - B_i(\widehat{\boldsymbol{\phi}}),$$

which is second-order unbiased estimator of R_{1i}, namely

$$E[\widehat{R_{1i}}(\widehat{\boldsymbol{\phi}})^{bc}] = R_{1i}(\boldsymbol{\phi}) + o(m^{-1}).$$

Now, we summarize the result for the second-order unbiased estimator of MSE in the following theorem.

Theorem 5.4. *Under Assumption 5.1, the second-order unbiased estimator of* MSE_i *is given by*

$$\widehat{\mathrm{MSE}}_i = \widehat{R_{1i}}(\widehat{\boldsymbol{\phi}})^{bc} + R_{2i}(\widehat{\boldsymbol{\phi}}) + 2R_{31i}(\widehat{\boldsymbol{\phi}}, \widehat{\boldsymbol{\kappa}}),$$

that is, $E\left[\widehat{\mathrm{MSE}}_i\right] = \mathrm{MSE}_i + o(m^{-1})$.

It is remarked that the proposed estimator of MSE does not require any resampling methods such as bootstrap. This means that the analytical estimator can be easily implemented and has less computational burden compared to bootstrap. Moreover, we do not assume normality of v_i and ε_{ij} in the derivation of the MSE estimator as in Lahiri and Rao (1995). Thus the proposed MSE estimator is expected to have a robustness property, which will be investigated in the simulation studies.

5.4 Numerical Studies

5.4.1 Model based simulation

We first compare the performances of EBLUP obtained from the proposed HNER with variance functions (HNERVF) with several existing models in terms of simulated mean squared errors (MSE). We consider the conventional nested error regression (NER) model, heteroscedastic NER model given by Jiang and Nguyen (2012) referred as JN, and the heteroscedastic NER

with random dispersions (HNERRD) proposed in Kubokawa et al. (2016). In applying the NER model, we use the unbiased estimator for variance components given in Prasad and Rao (1990) to calculate EBLUP. Further, we also consider the following log-link gamma mixed (GM) models as the competitor from the generalized linear mixed models, which also allows heteroscedasticity for the variances as the quadratic function of means. We used `glmer` function in `lme4` package in 'R' to apply the GM model.

In this simulation study, we set $m = 20$ and $n_i = 8$ in all cases, and we compute the simulated MSE in 10 scenarios denoted by S1, . . . ,S10. The simulated MSE for some area-specific parameter μ_i is define as

$$\mathrm{MSE}_i = \frac{1}{R}\sum_{r=1}^{R}(\widehat{\mu}_i^{(r)} - \mu_i^{(r)})^2, \tag{5.26}$$

where $R = 5000$ is the number of simulation runs, $\widehat{\mu}_i^{(r)}$ is the predicted value from some models and $\mu_i^{(r)}$ is the true values in the r-th iteration. In all scenarios, we generate covariates x_{ij}'s from the uniform distribution on $(0, 1)$, which are fixed in simulation runs. From S1 to S3, we consider the heteroscedastic model with area-level heteroscedastic variances given by

$$\mathrm{S1} \sim \mathrm{S3}: \quad y_{ij} = \beta_0 + \beta_1 x_{ij} + v_i + \varepsilon_{ij}, \quad v_i \sim (0, \tau^2), \quad \varepsilon_{ij} \sim (0, \sigma_i^2), \quad \mu_i = \beta_0 + v_i,$$

where $\sigma_i^2 = \exp(0.8 - z_i)$ and $(\beta_0, \beta_1, \tau) = (1, 0.5, 1.2)$. We generate z_i's from uniform distribution on $(-1, 1)$, which are fixed in simulation runs. The scenarios S1, S2 and S3 correspond to the cases where the distributions of both v_i and ε_{ij} are normal, t with 6 degrees of freedom, and chi-squared with 5 degrees of freedom, respectively, noting that both t-distribution and chi-squared distribution are scaled and located to meet the specified means and variances. For S4, we consider the homoscedastic model given by

$$\mathrm{S4}: \quad y_{ij} = \beta_0 + \beta_1 x_{ij} + v_i + \varepsilon_{ij}, \quad v_i \sim N(0, \tau^2), \quad \varepsilon_{ij} \sim N(0, \sigma^2), \quad \mu_i = \beta_0 + v_i,$$

with $(\beta_0, \beta_1, \tau, \sigma) = (1, 0.5, 1.2, 1.5)$. In S5 and S6, we use the heteroscedastic model with unit-level heteroscedastic variances given by

$$\mathrm{S5, S6}: \quad y_{ij} = \beta_0 + \beta_1 x_{ij} + v_i + \varepsilon_{ij}, \quad v_i \sim N(0, \tau^2), \quad \varepsilon_{ij} \sim N(0, \sigma_{ij}^2), \quad \mu_i = \beta_0 + v_i,$$

where $\sigma_{ij}^2 = \exp(0.8 - z_{ij})$ in S5 and $\sigma_{ij}^2 \sim \Gamma(5, 5/\exp(0.8 - z_{ij}))$ in S6. For S7 and S8, we consider the mixed model of the form

$$\mathrm{S7, S8}: \quad y_{ij} = \exp(\beta_0 + \beta_1 x_{ij} + v_i)\varepsilon_{ij}, \quad \mu_i = \exp(\beta_0 + v_i),$$

where $v_i \sim N(0, \tau^2)$, $\varepsilon_{ij} \sim \Gamma(3, 3)$ and $(\beta_0, \beta_1, \tau) = (0.5, 1, 0.3)$ in S7, and $v_i \sim t_6(0, \tau^2)$, $\varepsilon_{ij} \sim SLN(1, \sigma^2)$, and $(\beta_0, \beta_1, \tau, \sigma) = (1.2, 0.6, 0.4, 0.4)$ in S8, noting that $t_6(a, b)$ denotes the t-distribution with 6 degrees of freedom with mean a and variance b and $SLN(a, b)$ denotes the scaled log-normal distribution with mean a and variance b. Hence, $S7$ corresponds to the gamma mixed model with log-link function and $S8$ corresponds to its misspecified version. Finally, S9 to S10 are the mixed models defined as

$$\mathrm{S9}: \quad y_{ij} = (\beta_0 + \beta_1 x_{ij} + v_i)^2 \varepsilon_{ij}, \quad v_i \sim N(0, \tau^2), \quad \varepsilon_{ij} \sim SLN(1, \sigma^2), \quad \mu_i = (\beta_0 + v_i)^2$$

with $(\beta_0, \beta_1, \tau, \sigma) = (1, 0.6, 1.5, 0.5)$, and

$$\text{S10}: \quad y_{ij} = \{\exp(\beta_0 + \beta_1 x_{ij}) + v_i\}\varepsilon_{ij}, \quad v_i \sim N(0, \tau^2), \quad \varepsilon_{ij} \sim SLN(1, \sigma^2), \quad \mu_i = \exp(\beta_0) + v_i,$$

with $(\beta_0, \beta_1, \tau, \sigma) = (1, 0.3, 1.2, 0.5)$. It is noted that both S9 and S10 are also heteroscedastic model in the sense that $\text{Var}(y_{ij})$ depends on x_{ij}.

Under the 10 scenarios described above, we compute the simulated MSE values of predictors from five methods (HNERVF, HNERRD, NER, JN and GM) in each area. Since we can apply GM only to the data with positive y_{ij}'s, the MSE values of GM model are calculated from S7 to S10. In Table 8.1, we show the mean, max and min values of MSE over all areas for each model and scenario. From S1 to S3, it is observed that HNERVF performs better than the other models, and NER model performs worst since the true model is heteroscedastic. In S4, NER model performs best among four models since NER model is the true model and other HNER models are overfitted. It is also interesting to point out that the inefficiency of the prediction of JN is more serious than that of HNERVF and HNERRD. As in S5 and S6, the heteroscedastic variances are unit-level, the amount of improvement of HNERVF over other models gets greater. The scenario S7 corresponds to GM model, so that it is reasonable that MSE of GM is smallest among five models. The scenario S8 is not GM model but it is still close to GM model, in which GM model works well compared to the other models. However, once GM is seriously misspecified as in S9 and S10, GM does not work very much because of its somewhat strong parametric assumption. From S8 to S10, all models are misspecified, but HNERVF model works well compared to other models. Therefore, it is natural that HNERVF performs best when HNERVF is the true model, but even in case that HNERVF is misspecified, HNERVF also works reasonably well owing to its flexible structure of the model.

5.4.2 Finite sample performances of the MSE estimator

We next investigate the finite sample performances of the MSE estimators given in Theorem 5.4. To this end, we consider the data generating process given by

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + v_i + \varepsilon_{ij}, \quad v_i \sim (0, \tau^2), \quad \varepsilon_{ij} \sim (0, \exp(\gamma_0 + \gamma_1 z_{ij}))$$

with $\beta_0 = 1, \beta_1 = 0.8$, $\tau = 1.2$, $\gamma_0 = 1$ and $\gamma_1 = -0.4$. Moreover, we equally divided $m = 20$ areas into 5 groups ($G = 1, \ldots, 5$), so that each group has 4 areas and the areas in the same group has the same sample size $n_G = G+3$. Following Hall and Maiti (2006b), we consider five patterns of distributions of v_i and ε_{ij}, that is , M1: v_i and ε_{ij} are both normally distributed, M2: v_i and ε_{ij} are both scaled t-distribution with degrees of freedom 6, M3: v_i and ε_{ij} are both scaled and located χ_5 distribution, M4: v_i are ε_{ij} are scaled and located χ_5 and $-\chi_5$ distribution, respectively, and M5: v_i are ε_{ij} are both logistic distribution. The simulated values of the MSE are obtained from (5.26) based on $R = 10000$ simulation runs. Then, based on $R = 5000$ simulation runs, we calculate the relative bias (RB) and coefficient of variation (CV) of MSE estimators given by

$$\text{RB}_i = \frac{1}{R}\sum_{r=1}^{R} \frac{\widehat{\text{MSE}}_i^{(r)} - \text{MSE}_i}{\text{MSE}_i}, \qquad \text{CV}_i^2 = \frac{1}{R}\sum_{r=1}^{R} \left(\frac{\widehat{\text{MSE}}_i^{(r)} - \text{MSE}_i}{\text{MSE}_i} \right)^2$$

Table 5.1: Simulated Values of MSE for Various Scenarios and Models

	model	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
mean	HNERVF	0.368	0.370	0.371	0.311	0.280	0.293	0.269	0.619	0.198	0.376
	HNERRD	0.383	0.383	0.387	0.310	0.341	0.379	0.285	0.641	0.259	0.369
	NER	0.398	0.405	0.410	0.307	0.342	0.384	0.375	0.726	0.220	0.384
	JN	0.386	0.392	0.396	0.324	0.357	0.392	0.292	0.684	0.318	0.385
	GM	—	—	—	—	—	—	0.130	0.451	0.231	0.396
max	HNERVF	0.598	0.633	0.569	0.340	0.354	0.469	0.342	1.511	0.299	0.435
	HNERRD	0.630	0.634	0.603	0.342	0.424	0.523	0.405	1.603	0.415	0.419
	NER	0.642	0.639	0.596	0.339	0.423	0.526	0.518	1.992	0.336	0.439
	JN	0.634	0.643	0.618	0.372	0.445	0.545	0.426	1.834	0.532	0.441
	GM	—	—	—	—	—	—	0.149	0.970	0.372	0.473
min	HNERVF	0.138	0.145	0.150	0.272	0.202	0.196	0.205	0.398	0.142	0.297
	HNERRD	0.156	0.157	0.166	0.272	0.254	0.255	0.219	0.408	0.142	0.302
	NER	0.173	0.177	0.202	0.269	0.256	0.256	0.286	0.442	0.152	0.305
	JN	0.157	0.160	0.166	0.288	0.273	0.256	0.220	0.414	0.168	0.314
	GM	—	—	—	—	—	—	0.104	0.335	0.168	0.309

where $\widehat{\mathrm{MSE}}_i^{(r)}$ is the MSE estimator in the r-th iteration. In Table 5.2, we report mean and median values of RB_i and CV_i in each group. For comparison, results for the naive MSE estimator, without any bias correction, are reported in Table 5.2 as RBN. The naive MSE estimator is the plug-in estimator of the asymptotic MSE (5.17), namely it is obtained by replacing τ^2 and $\boldsymbol{\gamma}$ in formula (5.17) by $\widehat{\tau}^2$ and $\widehat{\boldsymbol{\gamma}}$, respectively. In Table 5.2, the relative bias is small, less than 10% in many cases. When the underlying distributions leave from normality, the MSE estimator still provides small relative bias although it has higher coefficient of variation. The naive MSE estimator is more biased than the analytical MSE estimator in all groups and models, so that the bias correction in MSE estimator is successful.

5.4.3 Real data application

We now apply the HNERVF model together with HNERRD, NER, JN and GM models considered in the simulation study in Section 5.4.1 to the data which originates from the posted land price (PLP) data along the Keikyu train line in 2001. This train line connects the suburbs in the Kanagawa prefecture to the Tokyo metropolitan area. Those who live in the suburbs in the Kanagawa prefecture take this line to work or study in Tokyo everyday, so that it is expected that the land price depends on the distance from Tokyo. The PLP data are available for 52 stations on the Keikyu train line, and we consider each station as a small area, namely, $m = 52$. For the i-th station, data of n_i land spots are available, where n_i varies around 4 and some areas have only one observation.

For $j = 1, \ldots, n_i$, y_{ij} denotes the scaled value of the PLP (Yen/10000) for the unit meter

Table 5.2: The Mean Values of Percentage Relative Bias (RB) and Coefficient of Variation (CV) of MSE Estimator and Relative Bias of Naive MSE Estimator (RBN) in Each Group.

Group	Measure	M1	M2	M3	M4	M5
G_1	RB	-8.72	-12.50	-10.86	-11.51	-11.81
	CV	17.48	23.60	23.47	23.40	21.24
	RBN	-12.67	-13.74	-13.10	-13.57	-13.39
G_2	RB	-7.61	-9.72	-10.58	-10.57	-7.27
	CV	17.52	23.24	22.70	23.03	20.31
	RBN	-10.16	-12.66	-11.48	-11.33	-10.54
G_3	RB	-7.89	-8.39	-7.65	-8.92	-6.34
	CV	19.85	26.05	24.66	25.37	22.94
	RBN	-9.31	-9.43	-8.70	-9.86	-7.58
G_4	RB	-6.52	-4.74	-4.96	-5.65	-4.27
	CV	22.02	28.37	26.93	27.68	24.98
	RBN	-10.83	-7.68	-7.98	-6.52	-6.42

squares of the j-th spot, T_i is the time to take from the nearby station i to the Tokyo station around 8:30 in the morning, D_{ij} is the value of geographical distance from the spot j to the station i and FAR_{ij} denotes the floor-area ratio, or ratio of building volume to lot area of the spot j. The three covariates FAR_{ij}, T_i and D_{ij} are also scaled by 100,10 and 1000, respectively. This data set is treated in Kubokawa et al. (2016), where they pointed out that the heteroscedasticity seem to be appropriate from boxplots of some areas and Bartlett test for testing homoscedastic variance. They used the PLP data with log-transformed observations, namely $\log y_{ij}$, but we use y_{ij} in this study since the results are easier to interpret than the results from $\log y_{ij}$. In the left panel of Figure 7.1, we show the plot of the pairs (D_{ij}, e_{ij}), where e_{ij} is OLS residuals defined as

$$e_{ij} = y_{ij} - (\widehat{\beta}_{0,OLS} + FAR_{ij}\widehat{\beta}_{1,OLS} + T_i\widehat{\beta}_{2,OLS} + D_{ij}\widehat{\beta}_{3,OLS}).$$

The figure indicates that the residuals are more variable for small D_{ij} than for large D_{ij}, and the variances are exponentially decreasing with respect to D_{ij}. Thus we apply the HNERVF model with the exponential variance function given by

$$y_{ij} = \beta_0 + FAR_{ij}\beta_1 + T_i\beta_2 + D_{ij}\beta_3 + v_i + \varepsilon_{ij}, \tag{5.27}$$

where $v_i \sim (0, \tau^2)$ and $\varepsilon_{ij} \sim (0, \exp(\gamma_0 + \gamma_1 D_{ij}))$. To compare the results, we also apply HNERRD, NER, JN and GM models to the PLP data with the same covariates. In applying NER model, we regard it as the submodel of HNERVF by putting $\gamma_1 = 0$ and use the same estimating method with HNERVF. The estimated regression coefficients from five models are given in the Table 5.3. We first note that the conditional expectation of the GM model is $\exp(\beta_0 + FAR_{ij}\beta_1 + T_i\beta_2 + D_{ij}\beta_3 + v_i)$, while that of other models has the liner form $\beta_0 + FAR_{ij}\beta_1 + T_i\beta_2 + D_{ij}\beta_3 + v_i$. Hence the scale of the estimated coefficients of GM are

different from those of other models. However, the signs of estimated coefficients are the same over all models. The resulting signs are intuitively natural since the PLP is expected to be decreasing as the distance between the spot and the nearest station gets large or the nearest station gets distant from Tokyo station. Moreover, in HNERVF model, the estimated value of γ_1 is $\widehat{\gamma}_1 = -1.82$, which is consistent to the observation from the left panel of Figure 7.1. Using the result of Theorem 5.1, the asymptotic standard error of $\widehat{\gamma}_1$ is 0.492, so that γ_1 seems significant.

We here consider to estimate the and price of a spot with floor-area ratio 100% and distance from 1000m from from the station i, namely $\mu_i = \beta_0 + \beta_1 + \beta_2 T_i + \beta_3 + v_i$ of HNERVF, HNERRD, NER and JN models, and $\mu_i = \exp(\beta_0 + \beta_1 + \beta_2 T_i + \beta_3 + v_i)$ of GM model. In Figure 5.2, we provide the predicted values of μ_i of each model. From the figure, we can observe that all five models provides relatively similar predicted values, and the predicted values tend to decrease with respect to the area index. This comes from the effect of T_i since T_i increase as the area index increases.

We finally calculate the mean squared errors (MSE) of predictors. In JN model, the consistent estimator of MSE cannot be obtained without any knowledge of grouping of areas (stations) as shown in Jiang and Nguyen (2012). For GM models, the second-order unbiased estimator of MSE is hard to obtain. Thus, we here consider the MSE estimator of HNERVF, HNERRD and NER models. We use the analytical estimator given in Theorem 5.4 for HNERVF and NER, and the parametric bootstrap MSE estimator developed in Kubokawa et al. (2016) is used for HNERRD with 1000 bootstrap replication. We found that the estimated MSE of HNERRD model is greater than 700 for all areas, while the estimated MSE of HNERVF and NER models are smaller than 20. The estimated value of shape parameter in dispersion (gamma) distribution in HNERRD is close to 2, which may inflate the MSE values. The estimated values of square root of MSE (RMSE) of HNERVF and NER models are given in the right panel of Figure 5.1. It is revealed that the estimated RMSE of HNERVF is smaller than that of NER in many areas. In particular, this is true in 37 areas among 52 areas. Especially, in the latter areas, it is observed that the amount of improvement is relatively large.

Table 5.3: The Estimated Regression Coefficients in Each Model

model	β_0	β_1	β_2	β_3
HNERVF	42.31	2.81	-3.56	-0.661
HNERRD	37.72	3.88	-3.24	-0.960
NER	33.35	6.58	-3.18	-0.832
JN	37.01	3.41	-2.59	-3.19
GM	3.63	0.168	-0.122	-0.039

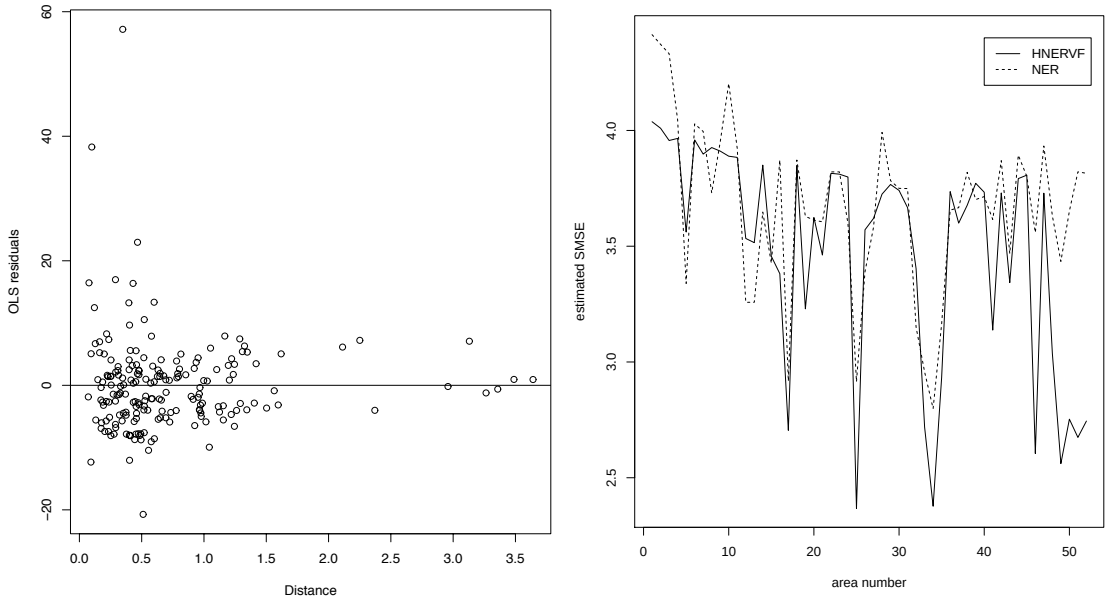

Figure 5.1: Scatter plot of OLS residuals against distance D_{ij} (left) and estimated square root of MSE (RMSE) in the HNERVF and NER models (right).

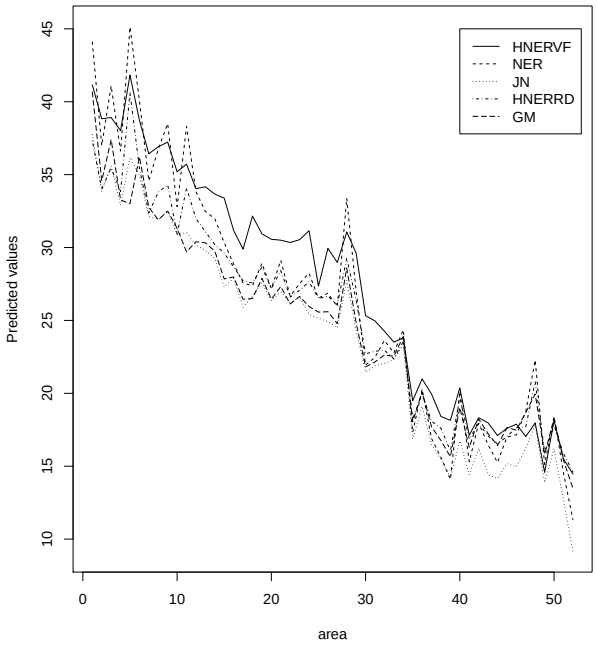

Figure 5.2: Predicted Values of μ_i in Each Model.

5.5 Technical Issues

5.5.1 Proof of Theorem 5.1

Since $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m$ are mutually independent, the consistency of $\widehat{\boldsymbol{\gamma}}$ follows from the standard argument, so that $\widehat{\tau}^2$ and $\widehat{\boldsymbol{\beta}}$ are also consistent. In what follows, we derive the asymptotic expressions of the estimators.

First we consider the asymptotic approximation of $\widehat{\tau}^2 - \tau^2$. From (5.8), we obtain

$$\begin{aligned}
\widehat{\tau}^2 - \tau^2 &= \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left\{(y_{ij} - \boldsymbol{x}_{ij}^t\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}})^2 - \widehat{\sigma}_{ij}^2\right\} - \tau^2 \\
&= \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left\{(y_{ij} - \boldsymbol{x}_{ij}^t\boldsymbol{\beta})^2 - \sigma_{ij}^2\right\} - \tau^2 - \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\sigma_{ij(1)}^2\boldsymbol{z}_{ij}^t(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) \\
&\quad - \frac{2}{N}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \boldsymbol{x}_{ij}^t\boldsymbol{\beta})\boldsymbol{x}_{ij}^t(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \boldsymbol{\beta}) + o_p(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) + o_p(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \boldsymbol{\beta}) \\
&= \frac{1}{m}\sum_{i=1}^{m}u_{1i} - \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\sigma_{ij(1)}^2\boldsymbol{z}_{ij}^t(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) + o_p(m^{-1/2}) + o_p(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}), \qquad (5.28)
\end{aligned}$$

where $u_{1i} = mN^{-1}\sum_{j=1}^{n_i}\left\{(y_{ij} - \boldsymbol{x}_{ij}^t\boldsymbol{\beta})^2 - \sigma_{ij}^2\right\} - \tau^2$ and we used the fact that $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \boldsymbol{\beta} = O_p(m^{-1/2})$ and $N^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \boldsymbol{x}_{ij}^t\boldsymbol{\beta})\boldsymbol{x}_{ij} = O_p(m^{-1/2})$ from the central limit theorem.

For the asymptotic expansion of $\widehat{\boldsymbol{\gamma}}$, remember that the estimator $\widehat{\boldsymbol{\gamma}}$ is given as the solution of the estimating equation

$$\frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left[\left\{y_{ij} - \bar{y}_i - (\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)^t\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}\right\}^2\boldsymbol{z}_{ij} - \sigma_{ij}^2(\boldsymbol{z}_{ij} - 2n_i^{-1}\boldsymbol{z}_{ij} + n_i^{-1}\bar{\boldsymbol{z}}_i)\right] = \boldsymbol{0}$$

Using Taylor expansions, we have

$$\begin{aligned}
\boldsymbol{0} &= \frac{1}{m}\sum_{i=1}\boldsymbol{u}_{2i} - \frac{2}{N}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left\{y_{ij} - \bar{y}_i - (\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)^t\boldsymbol{\beta}\right\}\boldsymbol{z}_{ij}(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)^t(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \boldsymbol{\beta}) \\
&\quad - \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\sigma_{ij(1)}^2(\boldsymbol{z}_{ij} - 2n_i^{-1}\boldsymbol{z}_{ij} + n_i^{-1}\bar{\boldsymbol{z}}_i)\boldsymbol{z}_{ij}^t(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) + o_p(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) + o_p(m^{-1/2}),
\end{aligned}$$

where

$$\boldsymbol{u}_{2i} = mN^{-1}\sum_{j=1}^{n_i}\left[\left\{y_{ij} - \bar{y}_i - (\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)^t\boldsymbol{\beta}\right\}^2\boldsymbol{z}_{ij} - \sigma_{ij}^2(\boldsymbol{z}_{ij} - 2n_i^{-1}\boldsymbol{z}_{ij} + n_i^{-1}\bar{\boldsymbol{z}}_i)\right].$$

From the central limit theorem, it follows that

$$\frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left\{y_{ij} - \bar{y}_i - (\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)^t\boldsymbol{\beta}\right\}\boldsymbol{z}_{ij}(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)^t = O_p(m^{-1/2}),$$

so that the second terms in the expansion formula is $o_p(m^{-1/2})$. Then we get

$$\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma} = \frac{N}{m} \left(\sum_{i=1}^{m} \sum_{j=1}^{n_i} \sigma^2_{ij(1)} (\boldsymbol{z}_{ij} - 2n_i^{-1} \boldsymbol{z}_{ij} + n_i^{-1} \bar{\boldsymbol{z}}_i) \boldsymbol{z}_{ij}^t \right)^{-1} \sum_{i=1}^{m} \boldsymbol{u}_{2i} + o_p(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) + o_p(m^{-1/2}).$$

Under Assumption 5.1, we have

$$\sum_{i=1}^{m} \sum_{j=1}^{n_i} \sigma^2_{ij(1)} (\boldsymbol{z}_{ij} - 2n_i^{-1} \boldsymbol{z}_{ij} + n_i^{-1} \bar{\boldsymbol{z}}_i) \boldsymbol{z}_{ij}^t = O(m).$$

From the independence of $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m$ and the fact $E(\boldsymbol{u}_{2i}) = \boldsymbol{0}$, we can use the central limit theorem to show that the leading term in the expansion of $\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}$ is $O_p(m^{-1/2})$. Thus,

$$\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma} = \frac{N}{m} \left(\sum_{i=1}^{m} \sum_{j=1}^{n_i} \sigma^2_{ij(1)} (\boldsymbol{z}_{ij} - 2n_i^{-1} \boldsymbol{z}_{ij} + n_i^{-1} \bar{\boldsymbol{z}}_i) \boldsymbol{z}_{ij}^t \right)^{-1} \sum_{i=1}^{m} \boldsymbol{u}_{2i} + o_p(m^{-1/2}).$$

Using the approximation of $\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}$ and $\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma} = O_p(m^{-1/2})$, we get the asymptotic expression of $\widehat{\tau}^2 - \tau^2$ from (5.28), which establishes the result for $\widehat{\tau}^2$ and $\widehat{\boldsymbol{\gamma}}$.

Finally we consider the asymptotic expansion of $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}$. From the expression in (5.6), it follows that

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta} + \sum_{s=1}^{q} \left(\frac{\partial}{\partial \gamma_s} \widetilde{\boldsymbol{\beta}} \right)^t (\widehat{\gamma}_s - \gamma) + \left(\frac{\partial}{\partial \tau^2} \widetilde{\boldsymbol{\beta}} \right)^t (\widehat{\tau}^2 - \tau^2) + o_p(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) + o_p(\widehat{\tau}^2 - \tau^2).$$

Since

$$\frac{\partial}{\partial \tau^2} \boldsymbol{\Sigma}_i = \boldsymbol{J}_{n_i}, \qquad \frac{\partial}{\partial \gamma_s} \boldsymbol{\Sigma}_i = \boldsymbol{W}_{i(s)}, \quad s = 1, \ldots, q,$$

for $\boldsymbol{W}_{i(s)} = \mathrm{diag}(\sigma^2_{i1(1)} z_{i1s}, \ldots, \sigma^2_{in_i(1)} z_{in_i s})$, we have

$$\begin{aligned} \frac{\partial}{\partial \tau^2} \widetilde{\boldsymbol{\beta}} &= \left(\boldsymbol{X}^t \boldsymbol{\Sigma}^{-1} \boldsymbol{X} \right)^{-1} \left(\sum_{i=1}^{m} \boldsymbol{X}_i^t \boldsymbol{\Sigma}_i^{-1} \boldsymbol{J}_{n_i} \boldsymbol{\Sigma}_i^{-1} \boldsymbol{X}_i \right) \left(\widetilde{\boldsymbol{\beta}}^*_\tau - \widetilde{\boldsymbol{\beta}} \right), \\ \frac{\partial}{\partial \gamma_s} \widetilde{\boldsymbol{\beta}} &= \left(\boldsymbol{X}^t \boldsymbol{\Sigma}^{-1} \boldsymbol{X} \right)^{-1} \left(\sum_{i=1}^{m} \boldsymbol{X}_i^t \boldsymbol{\Sigma}_i^{-1} \boldsymbol{W}_{i(s)} \boldsymbol{\Sigma}_i^{-1} \boldsymbol{X}_i \right) \left(\widetilde{\boldsymbol{\beta}}^*_{\gamma_s} - \widetilde{\boldsymbol{\beta}} \right), \quad s = 1 \ldots, q, \end{aligned} \tag{5.29}$$

where

$$\begin{aligned} \widetilde{\boldsymbol{\beta}}^*_\tau &= \left(\sum_{i=1}^{m} \boldsymbol{X}_i^t \boldsymbol{\Sigma}_i^{-1} \boldsymbol{J}_{n_i} \boldsymbol{\Sigma}_i^{-1} \boldsymbol{X}_i \right)^{-1} \sum_{i=1}^{m} \boldsymbol{X}_i^t \boldsymbol{\Sigma}_i^{-1} \boldsymbol{J}_{n_i} \boldsymbol{\Sigma}_i^{-1} \boldsymbol{y}_i, \\ \widetilde{\boldsymbol{\beta}}^*_{\gamma_s} &= \left(\sum_{i=1}^{m} \boldsymbol{X}_i^t \boldsymbol{\Sigma}_i^{-1} \boldsymbol{W}_{i(s)} \boldsymbol{\Sigma}_i^{-1} \boldsymbol{X}_i \right)^{-1} \sum_{i=1}^{m} \boldsymbol{X}_i^t \boldsymbol{\Sigma}_i^{-1} \boldsymbol{W}_{i(s)} \boldsymbol{\Sigma}_i^{-1} \boldsymbol{y}_i, \quad s = 1, \ldots, q. \end{aligned}$$

Under Assumption 5.1, we have $\widetilde{\boldsymbol{\beta}}^*_a - \boldsymbol{\beta} = O_p(m^{-1/2})$ for $a \in \{\tau, \gamma_1, \ldots, \gamma_q\}$, whereby $\widetilde{\boldsymbol{\beta}}^* - \widetilde{\boldsymbol{\beta}} = O_p(m^{-1/2})$. Since $\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma} = O_p(m^{-1/2})$ and $\widehat{\tau}^2 - \tau^2 = O_p(m^{-1/2})$ as shown above, we get

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \left(\boldsymbol{X}^t \boldsymbol{\Sigma}^{-1} \boldsymbol{X} \right)^{-1} \sum_{i=1}^{m} \boldsymbol{X}_i \boldsymbol{\Sigma}^{-1} (\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{\beta}) + o_p(m^{-1/2}),$$

which completes the proof.

5.5.2 Proof of Corollary 5.1

Let $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_{p+q+1})^t = (\boldsymbol{\beta}^t, \boldsymbol{\gamma}^t, \tau^2)^t$. Note that $\psi_i^{\phi_k}, k = 1, \ldots, p+q+1$ does not depend on $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{i-1}, \boldsymbol{y}_{i+1}, \ldots, \boldsymbol{y}_m$ and that $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m$ are mutually independent. Then,

$$\begin{aligned}\frac{1}{m^2} E\left[\left(\sum_{j=1}^m \psi_j^{\phi_k}\right)\left(\sum_{j=1}^m \psi_j^{\phi_l}\right)\middle|\boldsymbol{y}_i\right] &= \frac{1}{m^2}\sum_{j=1, j\neq i}^m E\left[\psi_j^{\phi_k}\psi_j^{\phi_l}\right] + \frac{1}{m^2}\psi_i^{\phi_k}\psi_i^{\phi_l} \\ &= \boldsymbol{\Omega}_{kl} + \frac{1}{m^2}\left\{\psi_i^{\phi_k}\psi_i^{\phi_l} - E\left[\psi_i^{\phi_k}\psi_i^{\phi_l}\right]\right\},\end{aligned}$$

where $\boldsymbol{\Omega}_{kl}$ is the (k,l)-element of $\boldsymbol{\Omega}$ and we used the fact that $E[\psi_j^{\phi_k}|\boldsymbol{y}_i] = E[\psi_j^{\phi_k}] = 0$ for $j \neq i$. Hence, we get the result from the asymptotic approximation of $\widehat{\boldsymbol{\phi}}$ given in Theorem 5.1.

5.5.3 Proof of Theorem 5.2.

We begin by deriving the conditional asymptotic bias of $\widehat{\boldsymbol{\gamma}}$. Let $\widetilde{\boldsymbol{\gamma}}$ be the solution of the equation

$$\boldsymbol{F}(\boldsymbol{\gamma};\boldsymbol{\beta}) \equiv \frac{1}{N}\sum_{i=1}^m\sum_{j=1}^{n_i}\left[\left\{y_{ij} - \bar{y}_i - (\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)^t\boldsymbol{\beta}\right\}^2 \boldsymbol{z}_{ij} - \sigma_{ij}^2(\boldsymbol{z}_{ij} - 2n_i^{-1}\boldsymbol{z}_{ij} + n_i^{-1}\bar{\boldsymbol{z}}_i)\right] = \boldsymbol{0}$$

with $\sigma_{ij}^2 = \sigma^2(\boldsymbol{z}_{ij}^t\boldsymbol{\gamma})$. For notational simplicity, we use $\boldsymbol{F}$ instead of $\boldsymbol{F}(\boldsymbol{\gamma};\boldsymbol{\beta})$ without any confusion and $F_r, r = 1, \ldots, q$ denotes the r-th component of $\boldsymbol{F}$, namely $\boldsymbol{F} = (F_1, \ldots, F_q)^t$. Define the derivatives $\boldsymbol{F}_{(\boldsymbol{a})}$ and $F_{h(\boldsymbol{ab})}$ by

$$\boldsymbol{F}_{(\boldsymbol{a})} = \frac{\partial \boldsymbol{F}}{\partial \boldsymbol{a}^t}, \qquad F_{r(\boldsymbol{ab})} = \frac{\partial^2 F_r}{\partial \boldsymbol{a}\partial \boldsymbol{b}^t}.$$

It is noted that $F_{h(\boldsymbol{\beta\gamma})} = 0$. Expanding $\boldsymbol{F}(\widehat{\boldsymbol{\gamma}}; \widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}) = \boldsymbol{0}$, we obtain

$$\boldsymbol{0} = \boldsymbol{F} + \boldsymbol{F}_{(\boldsymbol{\gamma})}(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) + \boldsymbol{F}_{(\boldsymbol{\beta})}(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \boldsymbol{\beta}) + \frac{1}{2}\boldsymbol{t}_1 + \frac{1}{2}\boldsymbol{t}_2 + o_p(m^{-1}),$$

where $\boldsymbol{t}_s = (t_{s1}, \ldots, t_{sq}), s = 1, 2$ for

$$t_{1r} = (\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})^t F_{r(\boldsymbol{\gamma\gamma})}(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}), \qquad t_{2r} = (\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \boldsymbol{\beta})^t F_{r(\boldsymbol{\beta\beta})}(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \boldsymbol{\beta}).$$

It is also noted that

$$\begin{aligned}\boldsymbol{F}_{(\boldsymbol{\gamma})} &= -\frac{1}{m}\sum_{k=1}^m\sum_{j=1}^{n_k}\sigma_{kj(1)}^2(\boldsymbol{z}_{kj} - 2n_k^{-1}\boldsymbol{z}_{kj} + n_k^{-1}\bar{\boldsymbol{z}}_k)\boldsymbol{z}_{kj}^t \\ \boldsymbol{F}_{(\boldsymbol{\beta})} &= -\frac{2}{N}\sum_{k=1}^m\sum_{j=1}^{n_k}\left\{y_{kj} - \bar{y}_k - (\boldsymbol{x}_{kj} - \bar{\boldsymbol{x}}_k)^t\boldsymbol{\beta}\right\}\boldsymbol{z}_{ij}(\boldsymbol{x}_{kj} - \bar{\boldsymbol{x}}_k)^t,\end{aligned}$$

so that $\boldsymbol{F}_{(\boldsymbol{\gamma})}$ is non-stochastic. Thus we have

$$E[\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}|\boldsymbol{y}_i] = -(\boldsymbol{F}_{(\boldsymbol{\gamma})})^{-1}\left\{E[\boldsymbol{F}(\boldsymbol{\gamma};\boldsymbol{\beta})|\boldsymbol{y}_i] + E\left[\boldsymbol{F}_{(\boldsymbol{\beta})}(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \boldsymbol{\beta})\middle|\boldsymbol{y}_i\right] + \frac{1}{2}E[\boldsymbol{t}_1|\boldsymbol{y}_i] + \frac{1}{2}E[\boldsymbol{t}_2|\boldsymbol{y}_i]\right\} + o_p(m^{-1}).$$

In what follows, we shall evaluate the each term in the parenthesis in the above expression. For the first term, since $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m$ are mutually independent and $E(\boldsymbol{u}_{2i}) = \boldsymbol{0}$, we have

$$E[\boldsymbol{F}(\boldsymbol{\gamma}; \boldsymbol{\beta})|\boldsymbol{y}_i] = \frac{1}{m}\boldsymbol{u}_{2i}.$$

For evaluation of the second term, we define $\boldsymbol{Z}_{kr} = \mathrm{diag}(z_{k1r}, \ldots, z_{kn_kr})$, where z_{kjr} denotes the r-th element of $\boldsymbol{z}_{kj}$. Then it follows that

$$\begin{aligned} E\left[\boldsymbol{F}_{r(\boldsymbol{\beta})}(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \boldsymbol{\beta})\Big|\boldsymbol{y}_i\right] &= -\frac{2}{N}\sum_{k=1}^{m} E\left[(\boldsymbol{y}_k - \boldsymbol{X}_k\boldsymbol{\beta})^t \mathrm{E}_k \boldsymbol{Z}_{kr} \mathrm{E}_k \boldsymbol{X}_k (\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \boldsymbol{\beta})\Big|\boldsymbol{y}_i\right] \\ &= -\frac{2}{N}\sum_{k=1, k\neq i}^{m} E\left[(\boldsymbol{y}_k - \boldsymbol{X}_k\boldsymbol{\beta})^t \mathrm{E}_k \boldsymbol{Z}_{kr} \mathrm{E}_k \boldsymbol{X}_k (\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \boldsymbol{\beta})\Big|\boldsymbol{y}_i\right] - \frac{2}{N}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^t \mathrm{E}_i \boldsymbol{Z}_{ir} \mathrm{E}_i \boldsymbol{X}_i E\left[\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \boldsymbol{\beta}\Big|\boldsymbol{y}_i\right]. \end{aligned}$$

Noting that it holds for $\ell = 1, \ldots, m$ and $k \neq i$

$$E\left[(\boldsymbol{y}_\ell - \boldsymbol{X}_\ell\boldsymbol{\beta})(\boldsymbol{y}_k - \boldsymbol{X}_k\boldsymbol{\beta})^t\Big|\boldsymbol{y}_i\right] = 1_{\{\ell=k\}}\boldsymbol{\Sigma}_k, \quad E[\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \boldsymbol{\beta}|\boldsymbol{y}_i] = \left(\boldsymbol{X}^t\boldsymbol{X}\right)^{-1}\boldsymbol{X}_i^t(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}),$$

we have

$$\begin{aligned} &E\left[(\boldsymbol{y}_k - \boldsymbol{X}_k\boldsymbol{\beta})^t \mathrm{E}_k \boldsymbol{Z}_{kr} \mathrm{E}_k \boldsymbol{X}_k (\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \boldsymbol{\beta})\Big|\boldsymbol{y}_i\right] \\ &\quad = \sum_{\ell=1}^{m} \mathrm{tr}\left\{\mathrm{E}_k \boldsymbol{Z}_{kr} \mathrm{E}_k \boldsymbol{X}_k (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}_k^t E\left[(\boldsymbol{y}_\ell - \boldsymbol{X}_\ell\boldsymbol{\beta})(\boldsymbol{y}_k - \boldsymbol{X}_k\boldsymbol{\beta})^t\Big|\boldsymbol{y}_i\right]\right\} \\ &\quad = \mathrm{tr}\left\{(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}_k^t\boldsymbol{\Sigma}_k \mathrm{E}_k \boldsymbol{Z}_{kr} \mathrm{E}_k \boldsymbol{X}_k\right\}, \end{aligned}$$

which is $O(m^{-1})$ and

$$\frac{1}{N}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^t \mathrm{E}_k \boldsymbol{Z}_{kr} \mathrm{E}_k \boldsymbol{X}_k E\left[\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \boldsymbol{\beta}\Big|\boldsymbol{y}_i\right] = o_p(m^{-1}).$$

Thus, we get

$$E\left[\boldsymbol{F}_{r(\boldsymbol{\beta})}(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \boldsymbol{\beta})\Big|\boldsymbol{y}_i\right] = -\frac{2}{m}\sum_{k=1}^{m}\sum_{j=1}^{n_k} \mathrm{tr}\left\{(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}_k^t\boldsymbol{\Sigma}_k \mathrm{E}_k \boldsymbol{Z}_{kr} \mathrm{E}_k \boldsymbol{X}_k\right\} + o_p(m^{-1}), \quad (5.30)$$

where the leading term is $O(m^{-1})$. For the third and forth terms, note that

$$F_{r(\boldsymbol{\gamma}\boldsymbol{\gamma})} = -\frac{1}{N}\sum_{k=1}^{m}\sum_{j=1}^{n_k} \sigma_{kj(2)}^2(\boldsymbol{z}_{kj} - 2n_k^{-1}\boldsymbol{z}_{kj} + n_k^{-1}\bar{\boldsymbol{z}}_k)\boldsymbol{z}_{kj}^t z_{kjr} \quad F_{r(\boldsymbol{\beta}\boldsymbol{\beta})} = \frac{2}{N}\sum_{k=1}^{m}\boldsymbol{X}_k^t \mathrm{E}_k \boldsymbol{Z}_{kr} \mathrm{E}_k \boldsymbol{X}_k,$$

which are non-stochastic. Then for $h = 1, \ldots, q$,

$$\begin{aligned} E[t_{1r}|\boldsymbol{y}_i] &= -\frac{1}{N}\sum_{k=1}^{m}\sum_{j=1}^{n_k} z_{kjr}\sigma_{kj(2)}^2(\boldsymbol{z}_{kj} - 2n_k^{-1}\boldsymbol{z}_{kj} + n_k^{-1}\bar{\boldsymbol{z}}_k)^t\boldsymbol{\Omega}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}\boldsymbol{z}_{kj} + o_p(m^{-1}), \\ E[t_{2r}|\boldsymbol{y}_i] &= \frac{2}{N}\sum_{k=1}^{m} \mathrm{tr}\left(\boldsymbol{X}_k^t \mathrm{E}_k \boldsymbol{Z}_{kr} \mathrm{E}_k \boldsymbol{X}_k \boldsymbol{V}_{\mathrm{OLS}}\right) + o_p(m^{-1}), \end{aligned}$$

for $\boldsymbol{V}_{\mathrm{OLS}} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{\Sigma}\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}$, where we used Corollary 5.1 and

$$E\left[(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \boldsymbol{\beta})(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \boldsymbol{\beta})^t \big| \boldsymbol{y}_i\right] = \boldsymbol{V}_{\mathrm{OLS}} + o_p(m^{-1}), \tag{5.31}$$

which follows from the similar argument to the proof of Corollary 5.1. Thus we obtain

$$E[\boldsymbol{t}_1|\boldsymbol{y}_i] = -\frac{1}{N}\sum_{k=1}^{m}\sum_{j=1}^{n_k} \boldsymbol{z}_{kj}\sigma^2_{kj(2)}(\boldsymbol{z}_{kj} - 2n_k^{-1}\boldsymbol{z}_{kj} + n_k^{-1}\bar{\boldsymbol{z}}_k)^t\boldsymbol{\Omega}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}\boldsymbol{z}_{kj} + o_p(m^{-1}),$$
$$E[\boldsymbol{t}_2|\boldsymbol{y}_i] = \frac{2}{N}\sum_{k=1}^{m}\left\{\mathrm{tr}\left(\boldsymbol{X}_k^t\mathrm{E}_k\boldsymbol{Z}_{kr}\mathrm{E}_k\boldsymbol{X}_k\boldsymbol{V}_{\mathrm{OLS}}\right)\right\}_r + o_p(m^{-1}),$$

where $\{\boldsymbol{a}_r\}_r$ denotes the q-dimensional vector $(a_1, \ldots, a_q)$. Therefore, we have established the result for $\widehat{\boldsymbol{\gamma}}$ in (5.15).

We next derive the result for $\widehat{\tau}^2$. Let

$$\widetilde{\tau}^2 = \frac{1}{N}\sum_{k=1}^{m}\left\{(\boldsymbol{y}_k - \boldsymbol{X}_k\boldsymbol{\beta})^t(\boldsymbol{y}_k - \boldsymbol{X}_k\boldsymbol{\beta}) - \sum_{j=1}^{n_k}\sigma^2_{kj}\right\}.$$

Using the Taylor series expansion, we have

$$\begin{aligned}\widehat{\tau}^2 = \widetilde{\tau}^2 &+ \frac{\partial\widetilde{\tau}^2}{\partial\boldsymbol{\gamma}}(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) + \frac{1}{2}(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})^t\left(\frac{\partial^2\widetilde{\tau}^2}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^t}\right)(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) \\ &+ \frac{\partial\widetilde{\tau}^2}{\partial\boldsymbol{\beta}}(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \boldsymbol{\beta}) + \frac{1}{2}(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \boldsymbol{\beta})^t\left(\frac{\partial^2\widetilde{\tau}^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^t}\right)(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \boldsymbol{\beta}) + o_p(m^{-1}),\end{aligned}$$

where we used the fact that $\partial^2\widetilde{\tau}^2/\partial\boldsymbol{\gamma}\partial\boldsymbol{\beta}^t = 0$. The straight calculation shows that

$$\frac{\partial\widetilde{\tau}^2}{\partial\boldsymbol{\gamma}} = -\frac{1}{N}\sum_{k=1}^{m}\sum_{j=1}^{n_k}\sigma^2_{kj(1)}\boldsymbol{z}_{kj}, \quad \frac{\partial^2\widetilde{\tau}^2}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^t} = -\frac{1}{N}\sum_{k=1}^{m}\sum_{j=1}^{n_k}\sigma^2_{kj(2)}\boldsymbol{z}_{kj}\boldsymbol{z}_{kj}^t, \quad \frac{\partial^2\widetilde{\tau}^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^t} = \frac{2}{N}\sum_{k=1}^{m}\boldsymbol{X}_i^t\boldsymbol{X}_i,$$

which are non-stochastic. Thus we obtain

$$\begin{aligned}E[\widehat{\tau}^2 - \tau^2|\boldsymbol{y}_i] &= E[\widetilde{\tau}^2 - \tau^2|\boldsymbol{y}_i] + \left(\frac{\partial\widetilde{\tau}^2}{\partial\boldsymbol{\gamma}}\right)^t E\left[\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}|\boldsymbol{y}_i\right] + \frac{1}{2}\mathrm{tr}\left\{\left(\frac{\partial^2\widetilde{\tau}^2}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^t}\right)E\left[(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})^t\big|\boldsymbol{y}_i\right]\right\} \\ &+ E\left[\left(\frac{\partial\widetilde{\tau}^2}{\partial\boldsymbol{\beta}}\right)^t(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \boldsymbol{\beta})\Big|\boldsymbol{y}_i\right] + \frac{1}{2}\mathrm{tr}\left\{\left(\frac{\partial^2\widetilde{\tau}^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^t}\right)E\left[(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \boldsymbol{\beta})(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \boldsymbol{\beta})^t\big|\boldsymbol{y}_i\right]\right\} + o_p(m^{-1}) \\ &\equiv B_{\tau 1}(\boldsymbol{y}_i) + B_{\tau 2}(\boldsymbol{y}_i) + B_{\tau 3}(\boldsymbol{y}_i) + B_{\tau 4}(\boldsymbol{y}_i) + B_{\tau 5}(\boldsymbol{y}_i) + o_p(m^{-1}).\end{aligned}$$

From the expression of $\widetilde{\tau}^2$, it holds that

$$\begin{aligned}B_{\tau 1}(\boldsymbol{y}_i) &= \frac{1}{N}\sum_{k=1,k\neq i}^{m} n_k\tau^2 + \frac{1}{N}\left\{(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^t(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}) - \sum_{j=1}^{n_i}\sigma^2_{ij}\right\} - \tau^2 \\ &= \left(1 - \frac{n_i}{N}\right)\tau^2 + \frac{1}{m}u_{1i} + \frac{n_i}{N}\tau^2 - \tau^2 = \frac{1}{m}u_{1i},\end{aligned}$$

for u_{1i} defined in (5.10). Also, we immediately have

$$B_{\tau 2}(\boldsymbol{y}_i) = -\frac{1}{N}\sum_{k=1}^{m}\sum_{j=1}^{n_k}\sigma^2_{kj(1)}\boldsymbol{z}^t_{kj}\boldsymbol{b}^{(i)}_{\gamma}(\boldsymbol{y}_i)$$

For evaluation of $B_{\tau 4}(\boldsymbol{y}_i)$, note that

$$\frac{\partial\widetilde{\tau}^2}{\partial\boldsymbol{\beta}} = -\frac{2}{N}\sum_{k=1}^{m}\boldsymbol{X}^t_k(\boldsymbol{y}_k - \boldsymbol{X}_k\boldsymbol{\beta}).$$

Similarly to (5.30), we get

$$\begin{aligned} B_{\tau 4}(\boldsymbol{y}_i) &= -\frac{2}{N}\sum_{k=1}^{m}E\left[(\boldsymbol{y}_k - \boldsymbol{X}_k\boldsymbol{\beta})^t\boldsymbol{X}_k(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \boldsymbol{\beta})\Big|\boldsymbol{y}_i\right] \\ &= -\frac{2}{N}\sum_{k=1}^{m}\mathrm{tr}\left\{(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t_k\boldsymbol{\Sigma}_k\boldsymbol{X}_k\right\} + o_p(m^{-1}). \end{aligned}$$

Moreover, Corollary 5.1 and (5.31) enable us to obtain the expression of $B_{\tau 3}(\boldsymbol{y}_i)$ and $B_{\tau 5}(\boldsymbol{y}_i)$, whereby we get

$$b^{(i)}_{\tau}(\boldsymbol{y}_i) = m^{-1}u_{1i} - \frac{1}{N}\sum_{k=1}^{m}\sum_{j=1}^{n_k}\sigma^2_{kj(1)}\boldsymbol{z}^t_{kj}\left\{\boldsymbol{b}^{(i)}_{\gamma}(\boldsymbol{y}_i) - \boldsymbol{b}_{\gamma}\right\} + b_{\tau},$$

which completes the proof for $\widehat{\tau}^2$ in (5.15).

We finally derive the result for $\widehat{\boldsymbol{\beta}}$. By the Taylor series expansion,

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta} + \sum_{s=1}^{q}\left(\frac{\partial}{\partial\gamma_s}\widetilde{\boldsymbol{\beta}}\right)(\widehat{\gamma}_s - \gamma) + \left(\frac{\partial}{\partial\tau^2}\widetilde{\boldsymbol{\beta}}\right)(\widehat{\tau}^2 - \tau^2) + o_p(m^{-1}),$$

since

$$\left(\frac{\partial\widetilde{\boldsymbol{\beta}}}{\partial\boldsymbol{\phi}}\right)^t(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi})(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi})^t\left(\frac{\partial\widetilde{\boldsymbol{\beta}}}{\partial\boldsymbol{\phi}}\right) = o_p(m^{-1}),$$

from $\partial\widetilde{\boldsymbol{\beta}}/\partial\boldsymbol{\phi} = O_p(m^{-1/2})$ as shown in the proof of Theorem 5.1. From (5.29), we have

$$\begin{aligned} &\sum_{s=1}^{q}\left(\frac{\partial}{\partial\gamma_s}\widetilde{\boldsymbol{\beta}}\right)(\widehat{\gamma}_s - \gamma_s) \\ &\quad= \left(\boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right)^{-1}\sum_{s=1}^{q}\left(\sum_{k=1}^{m}\boldsymbol{X}^t_i\boldsymbol{\Sigma}^{-1}_i\boldsymbol{W}_{i(s)}\boldsymbol{\Sigma}^{-1}_i\boldsymbol{X}_i\right)\left\{\left(\widetilde{\boldsymbol{\beta}}^*_{\gamma_s} - \boldsymbol{\beta}\right)(\widehat{\gamma}_s - \gamma_s) - (\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})(\widehat{\gamma}_s - \gamma_s)\right\}, \end{aligned}$$

and

$$\begin{aligned} &\left(\frac{\partial}{\partial\tau^2}\widetilde{\boldsymbol{\beta}}\right)(\widehat{\tau}^2 - \tau^2) \\ &\quad= \left(\boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right)^{-1}\left(\sum_{k=1}^{m}\boldsymbol{X}^t_k\boldsymbol{\Sigma}^{-1}_k\boldsymbol{J}_{n_k}\boldsymbol{\Sigma}^{-1}_k\boldsymbol{X}_k\right)\left\{(\widetilde{\boldsymbol{\beta}}^*_{\tau} - \boldsymbol{\beta})(\widehat{\tau}^2 - \tau^2) - (\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})(\widehat{\tau}^2 - \tau^2)\right\}. \end{aligned}$$

Let $\boldsymbol{\Omega}_{\beta^*\gamma_s} = E[(\widetilde{\boldsymbol{\beta}}^*_{\gamma_s} - \boldsymbol{\beta})(\widehat{\gamma}_s - \gamma_s)]$ and $\boldsymbol{\Omega}_{\beta^*\tau} = E[(\widetilde{\boldsymbol{\beta}}^*_{\tau} - \boldsymbol{\beta})(\widehat{\tau} - \tau)]$. Then it can be shown that

$$E[(\widetilde{\boldsymbol{\beta}}^*_{\tau} - \boldsymbol{\beta})(\widehat{\tau} - \tau)|\boldsymbol{y}_i] = \boldsymbol{\Omega}_{\beta^*\gamma_s} + o_p(m^{-1}), \quad E[(\widetilde{\boldsymbol{\beta}}^*_{\gamma_s} - \boldsymbol{\beta})(\widehat{\gamma}_s - \gamma_s)|\boldsymbol{y}_i] = \boldsymbol{\Omega}_{\beta^*\tau} + o_p(m^{-1}),$$

which can be proved by the same arguments as in Corollary 5.1. Thus from Corollary 5.1 and the fact that

$$E\left[\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}|\boldsymbol{y}_i\right] = \left(\boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}_i^t\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}),$$

we obtain the result for $\widehat{\boldsymbol{\beta}}$ in (5.15).

5.5.4 *Proof of (5.20).*

From the expansion of $\widehat{\mu}_i$, we have

$$E\left[(\widehat{\mu}_i - \widetilde{\mu}_i)^2\right] = E\left[\left\{\left(\frac{\partial\widetilde{\mu}_i}{\partial\boldsymbol{\phi}}\right)^t(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi})\right\}^2\right] + \frac{1}{2}U_1 + \frac{1}{4}U_2,$$

where

$$U_1 = E\left[\left(\frac{\partial\widetilde{\mu}_i}{\partial\boldsymbol{\phi}}\right)^t(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi})(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi})^t\left(\frac{\partial^2\widetilde{\mu}_i}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}^t}\Big|_{\boldsymbol{\phi}=\boldsymbol{\phi}^*}\right)(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi})\right]$$
$$U_2 = E\left[\left\{(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi})^t\left(\frac{\partial^2\widetilde{\mu}_i}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}^t}\Big|_{\boldsymbol{\phi}=\boldsymbol{\phi}^*}\right)(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi})\right\}^2\right].$$

It is noted that

$$U_1 = \sum_{j=1}^{p+q+1}\sum_{k=1}^{p+q+1}\sum_{\ell=1}^{p+q+1} E\left[\left(\frac{\partial\widetilde{\mu}_i}{\partial\phi_j}\right)\left(\frac{\partial^2\widetilde{\mu}_i}{\partial\phi_k\partial\phi_\ell}\Big|_{\boldsymbol{\phi}=\boldsymbol{\phi}^*}\right)(\widehat{\phi}_j - \phi_j)(\widehat{\phi}_k - \phi_k)(\widehat{\phi}_\ell - \phi_\ell)\right]$$
$$\equiv \sum_{j=1}^{p+q+1}\sum_{k=1}^{p+q+1}\sum_{\ell=1}^{p+q+1} U_{1jk\ell},$$

and

$$|U_{1jkl}| \le E\left[\left|\left(\frac{\partial\widetilde{\mu}_i}{\partial\phi_j}\right)\left(\frac{\partial^2\widetilde{\mu}_i}{\partial\phi_k\partial\phi_\ell}\Big|_{\boldsymbol{\phi}=\boldsymbol{\phi}^*}\right)\right|\left|(\widehat{\phi}_j - \phi_j)(\widehat{\phi}_k - \phi_k)(\widehat{\phi}_\ell - \phi_\ell)\right|\right]$$
$$\le E\left[\left|\left(\frac{\partial\widetilde{\mu}_i}{\partial\phi_j}\right)\left(\frac{\partial^2\widetilde{\mu}_i}{\partial\phi_k\partial\phi_\ell}\Big|_{\boldsymbol{\phi}=\boldsymbol{\phi}^*}\right)\right|^4\right]^{1/4} E\left[\left|(\widehat{\phi}_j - \phi_j)(\widehat{\phi}_k - \phi_k)(\widehat{\phi}_\ell - \phi_\ell)\right|^{4/3}\right]^{3/4} \quad (5.32)$$

using Holder's inequality. Since both $\partial\widetilde{\mu}_i/\partial\phi_j$ and $\partial^2\widetilde{\mu}_i/\partial\phi_k\partial\phi_\ell$ are linear functions of $\boldsymbol{y}_i$, the first term of (5.32) is finite under Assumption 5.1. Moreover, from Theorem 5.1, it follows $\sqrt{m}|\widehat{\phi}_j - \phi_j| \le C(\boldsymbol{y})$ for some quadratic function of $\boldsymbol{y}$, so that the second term in (5.32) is also finite. Hence, we have $U_1 = o(m^{-1})$. Similarly, we also obtain $U_2 = o(m^{-1})$. Therefore,

using Corollary 5.1, we have

$$
\begin{aligned}
E\left[(\widehat{\mu}_i-\widetilde{\mu}_i)^2\right] &= E\left[\left\{\left(\frac{\partial\widetilde{\mu}_i}{\partial\boldsymbol{\phi}}\right)^t(\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi})\right\}^2\right]+o(m^{-1})\\
&= \operatorname{tr}\left\{E\left[\left(\frac{\partial\widetilde{\mu}_i}{\partial\boldsymbol{\phi}}\right)\left(\frac{\partial\widetilde{\mu}_i}{\partial\boldsymbol{\phi}}\right)^t E\left((\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi})(\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi})^t\Big|\boldsymbol{y}_i\right)\right]\right\}+o(m^{-1})\\
&= \operatorname{tr}\left\{E\left[\left(\frac{\partial\widetilde{\mu}_i}{\partial\boldsymbol{\phi}}\right)\left(\frac{\partial\widetilde{\mu}_i}{\partial\boldsymbol{\phi}}\right)^t\boldsymbol{\Omega}+\left(\frac{\partial\widetilde{\mu}_i}{\partial\boldsymbol{\phi}}\right)\left(\frac{\partial\widetilde{\mu}_i}{\partial\boldsymbol{\phi}}\right)^t c(\boldsymbol{y}_i)o(m^{-1})\right]\right\}+o(m^{-1})\\
&= \operatorname{tr}\left\{E\left[\left(\frac{\partial\widetilde{\mu}_i}{\partial\boldsymbol{\phi}}\right)\left(\frac{\partial\widetilde{\mu}_i}{\partial\boldsymbol{\phi}}\right)^t\right]\boldsymbol{\Omega}\right\}+o(m^{-1})
\end{aligned}
$$

since $c(\boldsymbol{y}_i)$ is fourth-order function of $\boldsymbol{y}_i$ and $\partial\widetilde{\mu}_i/\partial\boldsymbol{\phi}$ is a linear function of $\boldsymbol{y}_i$, which completes the proof.

5.5.5 *Derivation of* $R_{31i}(\boldsymbol{\phi},\boldsymbol{\kappa})$.

Since $\boldsymbol{y}_i$ given $v_i,\boldsymbol{\epsilon}_i$ is non-stochastic, we have

$$
\begin{aligned}
&E\left[\left(\frac{\partial\widetilde{\mu}_i}{\partial\boldsymbol{\phi}}\right)^t(\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi})w_i\right]\\
&= E\left[E\left[\left(\frac{\partial\widetilde{\mu}_i}{\partial\boldsymbol{\phi}}\right)^t(\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi})w_i\Big|v_i,\boldsymbol{\epsilon}_i\right]\right]=E\left[E(\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi}|\boldsymbol{y}_i)^t\left(\frac{\partial\widetilde{\mu}_i}{\partial\boldsymbol{\phi}}\right)w_i\right]\\
&= E\left[\boldsymbol{b}_{\boldsymbol{\beta}}^{(i)}(\boldsymbol{y}_i)^t\left(\frac{\partial\widetilde{\mu}_i}{\partial\boldsymbol{\beta}}\right)w_i\right]+E\left[\boldsymbol{b}_{\boldsymbol{\gamma}}^{(i)}(\boldsymbol{y}_i)^t\left(\frac{\partial\widetilde{\mu}_i}{\partial\boldsymbol{\gamma}}\right)w_i\right]+E\left[b_{\tau}^{(i)}(\boldsymbol{y}_i)\left(\frac{\partial\widetilde{\mu}_i}{\partial\tau}\right)w_i\right]+o(m^{-1})\\
&\equiv R_{31i}(\boldsymbol{\phi})+o(m^{-1}).
\end{aligned}
$$

It is noted that $E(w_i)=0$ and

$$
E\left[(y_{ij}-\boldsymbol{x}_{ij}^t\boldsymbol{\beta})w_i\right]=E\left[(v_i+\varepsilon_{ij})w_i\right]=\left(\sum_{j=1}^{n_i}\lambda_{ij}-1\right)\tau^2+\sum_{j=1}^{n_i}\lambda_{ij}\sigma_{ij}^2=0. \tag{5.33}
$$

Using the expression (5.15) and (5.19), it follows that

$$
\begin{aligned}
E\left[\boldsymbol{b}_{\boldsymbol{\beta}}^{(i)}(\boldsymbol{y}_i)^t\left(\frac{\partial\widetilde{\mu}_i}{\partial\boldsymbol{\beta}}\right)w_i\right] &= \left(\boldsymbol{c}_i-\sum_{j=1}^{n_i}\lambda_{ij}\boldsymbol{x}_{ij}\right)^t\left(\boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}_i^t\boldsymbol{\Sigma}_i^{-1}E\left[(\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta})w_i\right]=0\\
E\left[\boldsymbol{b}_{\boldsymbol{\gamma}}^{(i)}(\boldsymbol{y}_i)^t\left(\frac{\partial\widetilde{\mu}_i}{\partial\boldsymbol{\gamma}}\right)w_i\right] &= \eta_i^{-2}\sum_{j=1}^{n_i}\sigma_{ij}^{-2}\boldsymbol{\delta}_{ij}^t\left(\sum_{k=1}^{m}\sum_{h=1}^{n_k}\sigma_{kh(1)}^2\boldsymbol{z}_{kh}\boldsymbol{z}_{kh}^t\right)^{-1}\boldsymbol{M}_{2ij}(\boldsymbol{\phi},\boldsymbol{\kappa})\\
E\left[b_{\tau}^{(i)}(\boldsymbol{y}_i)\left(\frac{\partial\widetilde{\mu}_i}{\partial\tau}\right)w_i\right] &= m^{-1}\eta_i^{-2}\sum_{j=1}^{n_i}\sigma_{ij}^{-2}\Big\{M_{1ij}(\boldsymbol{\phi},\boldsymbol{\kappa})-\boldsymbol{T}_1(\boldsymbol{\gamma})^t\boldsymbol{T}_2(\boldsymbol{\gamma})\boldsymbol{M}_{2ij}(\boldsymbol{\phi},\boldsymbol{\kappa})\Big\},
\end{aligned}
$$

where

$$\boldsymbol{M}_{2ij}(\boldsymbol{\phi},\boldsymbol{\kappa}) = E\left[\boldsymbol{u}_{2i}(y_{ij} - \boldsymbol{x}_{ij}^t\boldsymbol{\beta})w_i\right], \quad M_{1ij}(\boldsymbol{\phi},\boldsymbol{\kappa}) = E\left[u_{1i}(y_{ij} - \boldsymbol{x}_{ij}^t\boldsymbol{\beta})w_i\right].$$

To evaluate M_{1ij} and $\boldsymbol{M}_{2ij}$, we first prove the following result for fixed $j,k,\ell \in \{1,\ldots,n_i\}$.

$$E\big[(v_i+\varepsilon_{ij})(v_i+\varepsilon_{ik})(v_i+\varepsilon_{i\ell})w_i\big] = \tau^2\eta_i^{-1}\Big[\tau^2(3-\kappa_v) + \kappa_\varepsilon\sigma_{ij}^2 1_{\{j=k=\ell\}} + \sigma_{ij}^2(1_{\{j=k\neq\ell\}} - 1_{\{j=k\}}) + \sigma_{ij}^2(1_{\{j=\ell\neq k\}} - 1_{\{j=\ell\}}) + \sigma_{ik}^2(1_{\{k=\ell\neq j\}} - 1_{\{k=\ell\}})\Big]. \tag{5.34}$$

To show (5.34), we note that the left side can be rewritten as

$$-\eta_i^{-1}E\left[(v_i+\varepsilon_{ij})(v_i+\varepsilon_{ik})(v_i+\varepsilon_{i\ell})v_i\right] + \sum_{h=1}^{n_i}\lambda_{ih}E\left[(v_i+\varepsilon_{ij})(v_i+\varepsilon_{ik})(v_i+\varepsilon_{i\ell})\varepsilon_{ih}\right] \tag{5.35}$$

from the definition of w_i. Using the fact that $\varepsilon_{i1},\ldots,\varepsilon_{in_i}$ and v_i are independent, the first term in (5.35) is calculated as

$$E\left[v_i^4 + (\varepsilon_{ij}\varepsilon_{ik} + \varepsilon_{ij}\varepsilon_{i\ell} + \varepsilon_{ik}\varepsilon_{i\ell})v_i^2\right] = \kappa_v\tau^4 + \tau^2\left(\sigma_{ij}^2 1_{\{j=k\}} + \sigma_{ij}^2 1_{\{j=\ell\}} + \sigma_{ik}^2 1_{\{k=\ell\}}\right).$$

Moreover, we have

$$\begin{aligned} E\left[(v_i+\varepsilon_{ij})(v_i+\varepsilon_{ik})(v_i+\varepsilon_{i\ell})\varepsilon_{ih}\right] &= E\left[\varepsilon_{ih}(\varepsilon_{ij}+\varepsilon_{i\ell}+\varepsilon_{ik})v_i^2 + \varepsilon_{ij}\varepsilon_{ik}\varepsilon_{i\ell}\varepsilon_{ih}\right] \\ &= \tau^2\sigma_{ih}^2\left(1_{\{h=j\}} + 1_{\{h=k\}} + 1_{\{h=\ell\}}\right) + \kappa_\varepsilon\sigma_{ih}^4 1_{\{j=k=\ell=h\}} \\ &\quad + \sigma_{ih}^2\left(\sigma_{ij}^2 1_{\{j=k\neq\ell=h\}} + \sigma_{ij}^2 1_{\{j=\ell\neq k=h\}} + \sigma_{ik}^2 1_{\{j=h\neq k=\ell\}}\right), \end{aligned}$$

whereby the second term in (5.35) can be calculated as

$$\tau^2\eta_i^{-1}\left[3\tau^2 + \kappa_\varepsilon\sigma_{ij}^2 1_{\{j=k=\ell\}} + \sigma_{ij}^2 1_{\{j=k\neq\ell\}} + \sigma_{ij}^2 1_{\{j=\ell\neq k\}} + \sigma_{ik}^2 1_{\{k=\ell\neq j\}}\right],$$

where we used the expression $\lambda_{ih} = \tau^2\eta_i^{-1}\sigma_{ih}^{-2}$. Then we established the result (5.34). From (5.34), we immediately have

$$\begin{aligned} \sum_{\ell=1}^{n_i} E\big[(v_i+\varepsilon_{ij})(v_i+\varepsilon_{ik})(v_i+\varepsilon_{i\ell})w_i\big] &= \tau^2\eta_i^{-1}\left[n_i\tau^2(3-\kappa_v) + \sigma_{ij}^2(\kappa_\varepsilon - 3)1_{\{j=k\}}\right] \\ &= E\big[(v_i+\varepsilon_{ij})(v_i+\varepsilon_{ik})^2 w_i\big]. \end{aligned}$$

Now, we return to the evaluation of M_{1ij} and $\boldsymbol{M}_{2ij}$. It follows that

$$\begin{aligned} M_{1ij}(\boldsymbol{\phi},\boldsymbol{\kappa}) &= \frac{m}{N}\sum_{h=1}^{n_i} E\left[(y_{ih} - \boldsymbol{x}_{ih}^t\boldsymbol{\beta})^2(y_{ij} - \boldsymbol{x}_{ij}^t\boldsymbol{\beta})w_i\right] \\ &= mN^{-1}\eta_i^{-1}\tau^2\Big\{n_i\tau^2(3-\kappa_v) + \sigma_{ij}^2(\kappa_\varepsilon - 3)\Big\} \end{aligned}$$

and

$$\begin{aligned}\boldsymbol{M}_{2ij}(\boldsymbol{\phi},\boldsymbol{\kappa}) &= \frac{m}{N}\sum_{h=1}^{n_i}\boldsymbol{z}_{ih}E\left[\{v_i+\varepsilon_{ih}-(v_i+\bar{\varepsilon}_i)\}^2(v_i+\varepsilon_{ij})w_i\right] \\ &= \frac{m}{N}\sum_{h=1}^{n_i}\boldsymbol{z}_{ih}\Big\{E\left[(v_i+\varepsilon_{ih})^2(v_i+\varepsilon_{ij})w_i\right]-2n_i^{-1}\sum_{k=1}^{n_i}E\left[(v_i+\varepsilon_{ij})(v_i+\varepsilon_{ik})(v_i+\varepsilon_{ih})w_i\right] \\ &\quad + n_i^{-2}\sum_{k=1}^{n_i}\sum_{\ell=1}^{n_i}E\left[(v_i+\varepsilon_{ij})(v_i+\varepsilon_{ik})(v_i+\varepsilon_{i\ell})w_i\right]\Big\}.\end{aligned}$$

Using the identity given in (5.34), we have

$$\begin{aligned}\boldsymbol{M}_{2ij}(\boldsymbol{\phi},\boldsymbol{\kappa}) &= mN^{-1}\tau^2\eta_i^{-1}\sum_{h=1}^{n_i}\boldsymbol{z}_{ih}\Big\{\sigma_{ij}^2(\kappa_\varepsilon-3)(1_{\{j=h\}}-2n_i^{-1}1_{\{j=h\}}+n_i^{-2})\Big\} \\ &= mN^{-1}\tau^2\eta_i^{-1}n_i^{-2}(n_i-1)^2(\kappa_\varepsilon-3)\sigma_{ij}^2\boldsymbol{z}_{ij},\end{aligned}$$

which completes the result in (5.22).

5.5.6 Evaluation of $R_{32i}(\boldsymbol{\phi})$.

Since $\boldsymbol{y}_i$ given v_i and $\boldsymbol{\epsilon}_i$ is non-stochastic, we have

$$\begin{aligned}R_{32i}(\boldsymbol{\phi}) &= \frac{1}{2}E\left[(\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi})^t\left(\frac{\partial^2\widetilde{\mu}_i}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}^t}\Big|_{\boldsymbol{\phi}=\boldsymbol{\phi}^*}\right)(\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi})w_i\right] \\ &= \frac{1}{2}E\left[E\left[(\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi})^t\left(\frac{\partial^2\widetilde{\mu}_i}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}^t}\Big|_{\boldsymbol{\phi}=\boldsymbol{\phi}^*}\right)(\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi})w_i\Big|v_i,\boldsymbol{\epsilon}_i\right]\right] \\ &= \frac{1}{2}\mathrm{tr}\left\{\boldsymbol{\Omega}E\left[\left(\frac{\partial^2\widetilde{\mu}_i}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}^t}\Big|_{\boldsymbol{\phi}=\boldsymbol{\phi}^*}\right)w_i\right]\right\}+o(m^{-1})E\left[\mathrm{tr}\left\{c(\boldsymbol{y}_i)\left(\frac{\partial^2\widetilde{\mu}_i}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}^t}\Big|_{\boldsymbol{\phi}=\boldsymbol{\phi}^*}\right)\right\}w_i\right],\end{aligned}$$

where we used Corollary 5.1 in the last equation. Note that

$$\frac{\partial^2\widetilde{\mu}_i}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}^t}\Big|_{\boldsymbol{\phi}=\boldsymbol{\phi}^*}=\frac{\partial^2\widetilde{\mu}_i}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}^t}+\sum_{k=1}^{p+q+1}(\phi_k^*-\phi_k)\left(\frac{\partial^3\widetilde{\mu}_i}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}^t\partial\phi_k}\Big|_{\phi_k=\phi_k^{**}}\right), \tag{5.36}$$

where ϕ_k^{**} is an intermediate value between ϕ_k^* and ϕ_k. Further note that the third order partial derivatives of $\widetilde{\mu}_i$ is a linear function of $\boldsymbol{y}_i$, so that the second term of R_{32i} is $o(m^{-1})$. Similarly, it follows that

$$E\left[\left(\frac{\partial^2\widetilde{\mu}_i}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}^t}\Big|_{\boldsymbol{\phi}=\boldsymbol{\phi}^*}\right)w_i\right]=E\left[\left(\frac{\partial^2\widetilde{\mu}_i}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}^t}\right)w_i\right]+o(1)=o(1),$$

since the second order partial derivatives of $\widetilde{\mu}_i$ is a linear function of $y_{ij}-\boldsymbol{x}_{ij}^t\boldsymbol{\beta}$ and the identity (5.33). Therefore, we finally get $R_{32i}(\boldsymbol{\phi})=o(m^{-1})$.

Chapter 6

Shrinking Both Means and Variances

6.1 Introduction

In the Fay-Herriot model (2.1), it is conventionally assumed that the sampling variances D_i are known. In practice, however, the sampling variances are often estimated in various ways, and the small area estimators are provided by replacing the known variances with their estimators. This means that the small area estimators derived in the Fay-Herriot model involve substantial errors which come from estimation of variance, and we need to evaluate the estimation errors. To this end, several approaches are developed in the small area literature, for example, Arora and Lahiri (1997), You and Chapman (2006), and Wang and Fuller (2003).

You and Chapman (2006) proposed the modified Fay-Herriot model taking the estimated sampling variance into the Fay-Herriot model. To describe their model, suppose that there are m small areas, and let (X_i, S_i^2) be a pair of direct survey estimates of mean and variance in the i-th small area for $i = 1, \ldots, m$. Let $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{ip})^t$ be a vector of p covariates available at the estimation stage. Then the Fay-Herriot model can be modified as

$$\begin{aligned} &X_i|\theta_i, \sigma_i^2 \sim N(\theta_i, \sigma_i^2), \quad \theta_i \sim N(\boldsymbol{z}_i^t\boldsymbol{\beta}, \tau^2) \\ &S_i^2|\sigma_i^2 \sim \Gamma\left(\frac{n_i-1}{2}, \frac{n_i-1}{2\sigma_i^2}\right), \quad \sigma_i^2 \sim \pi(\sigma_i^2) \end{aligned} \tag{6.1}$$

where $(X_i, S_i^2, \theta_i, \sigma_i^2)$, $i = 1, \ldots, m$, are mutually independent and $\Gamma(a, b)$ denotes the gamma distribution with density proportional to $x^{\alpha-1}\exp(-\beta x)$, $x > 0$. Here, n_i is the sample size for a simple random sample in the i-th area, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^t$ is the $p \times 1$ vector of regression coefficients. In the framework of (6.1), You and Chapman (2006) suggested the hierarchical Bayesian approach by setting prior distributions:

$$\pi(\boldsymbol{\beta}) \propto 1, \quad \sigma_i^2 \sim IG(a_i, b_i), \ i = 1, \ldots, m, \quad \tau^2 \sim IG(a_0, b_0),$$

where $IG(a, b)$ is the inverse Gamma density function with density proportional to $x^{-\alpha-1}\exp(-\beta/x)$, $x > 0$, and a_i, b_i $(i = 0, \ldots, m)$ are chosen to be very small known constants, so that the prior distributions on σ_i^2 and τ^2 are close to the uniform distribution. However, the nearly uniform prior distribution for σ_i^2 does not produce shrinkage estimation of the sampling variances.

On the other hand, recently, Maiti et al. (2014) proposed the empirical Bayes approach for (6.1), namely

$$\begin{aligned} &X_i|\theta_i,\sigma_i^2 \sim N(\theta_i,\sigma_i^2), \quad \theta_i \sim N(\boldsymbol{z}_i^t\boldsymbol{\beta},\tau^2) \\ &S_i^2|\sigma_i^2 \sim \Gamma\left(\frac{n_i-1}{2},\frac{n_i-1}{2\sigma_i^2}\right), \quad \sigma_i^2 \sim IG(\alpha,\gamma), \end{aligned} \tag{6.2}$$

where $\boldsymbol{\beta},\tau^2,\alpha$ and γ are unknown parameters. They estimated model parameters $\boldsymbol{\beta}$ and τ^2 as well as α and γ from the (marginal) likelihood function. However, the marginal likelihood function cannot be obtained in a closed form and they developed the EM algorithm for getting estimates of the model parameters. Also we found through the simulation study that the estimates of (γ,α) tend to be unstable. Moreover, the analytical expression of the Bayes estimator of θ_i is hard to obtain since the posterior distribution of θ_i is no longer a normal distribution but an unfamiliar distribution. Thus, it is worth developing much easier yet practical method shrinking both means and variances in small area estimation.

These observations motivate us to propose the Bayesian approach for small area models shrinking both mean and variances. To achieve this, we assume the uniform prior distributions on τ^2 and $\boldsymbol{\beta}$, namely $\pi(\boldsymbol{\beta},\tau^2) \propto 1$, and the following structure is introduced for σ_i^2:

$$\sigma_i^2 \sim IG(a_i,b_i\gamma), \quad i=1,\ldots,m, \quad \pi(\gamma) \propto 1,$$

where a_i and b_i are constants specified by users. A suggestion for the choice of a_i and b_i is given in the end of Section 6.2.1. In these settings, the full conditional posterior distributions are all familiar forms that enable us to easily draw the samples via the Markov chain Monte Carlo technique, in particular the Gibbs sampler as discussed in Section 6.2. Using these posterior samples, we obtain the point estimates of the parameter of interest θ_i by the simple average of posterior samples. Moreover, the prediction intervals are easily constructed from quantiles of posterior samples compared to the empirical Bayes confidence intervals given in Dass et al. (2012) and Hwang et al. (2009). In Section 6.2.2, we also consider the alternative formulation of the true variance σ_i^2 in each area with use of covariate information, namely σ_i^2 is structured as $\sigma_i^2 \sim IG(a_i, b_i\gamma\exp(\boldsymbol{w}_i^t\boldsymbol{\eta}))$ for some vector of covariates $\boldsymbol{w}_i$ and unknown regression vector of coefficients $\boldsymbol{\eta}$. In this paper, we also develop a Bayesian method for this model and prove the posterior propriety and finiteness of the posterior variances when we use the improper priors for unknown parameters.

This chapter is organized as follows: In Section 6.2, the full Bayesian model alternative to Maiti et al. (2014) and You and Chapman (2006) is proposed. The full conditional distribution is described, and the Gibbs sampling for MCMC is given. As a theoretical main result, under a mild sufficient condition, we prove that the resulting posterior distribution is proper and the model parameters have finite variances. In Section 6.3, we carry out simulation studies to compare the suggested methods with the models by Maiti et al. (2014) and You and Chapman (2006). As real data analysis, we apply our methods to two real data sets, the SFIE data in Japan and the famous corn crop data, in Section 6.4. The proofs of the main theorem are given in Section 6.5.

6.2 Bayesian models shrinking both means and variances

6.2.1 Model settings and Bayesian inferences

We propose Bayesian multi-stage small area model shrinking both means and variances described as

$$\begin{aligned} &X_i|\theta_i,\sigma_i^2 \sim N(\theta_i,\sigma_i^2), \quad \theta_i|\boldsymbol{\beta},\tau^2 \sim N(\boldsymbol{z}_i^t\boldsymbol{\beta},\tau^2), \\ &S_i^2|\sigma_i^2 \sim \Gamma\left(\frac{n_i-1}{2},\frac{n_i-1}{2\sigma_i^2}\right), \quad \sigma_i^2|\gamma \sim IG(a_i,b_i\gamma) \\ &\pi(\boldsymbol{\beta},\tau^2,\gamma)=1, \end{aligned} \tag{6.3}$$

where $(X_i, S_i^2, \theta_i, \sigma_i^2)$, $i = 1, \ldots, m$, are conditionally independent given $(\boldsymbol{\beta}, \tau^2, \gamma)$. Here, a_i, b_i are positive and known (user specified) constants. The choice of a_i and b_i is not concerned with the propriety of the posterior distributions given in Theorem 6.1 as far as a_i and b_i are positive. The practical choice of these constants is discussed later. Note that the model for S_i^2 in (6.3) means that $(n_i - 1)S_i^2/\sigma_i^2$ given σ_i^2 follows a chi-square distribution with $(n_i - 1)$ degrees of freedom. This setting is appropriate under simple random sampling, but for complex sampling design, the degrees of freedom needs to be determined carefully as discussed in Maples et al. (2009).

We now consider the posterior distribution and investigate its properties. We denote $D = \{X_i, S_i^2, \boldsymbol{z}_i\}_{i=1,\ldots,m}$, the set of all observed data, for notational simplicity. From the formulation (6.3), the posterior density is given by

$$\begin{aligned} &\pi(\theta_1,\ldots,\theta_m,\sigma_1^2,\ldots,\sigma_m^2,\boldsymbol{\beta},\tau^2,\gamma|D) \\ &\quad\propto (\tau^2)^{-m/2}\prod_{i=1}^m \gamma^{a_i}(\sigma_i^2)^{-n_i/2-a_i-1}\exp\left\{-\frac{(X_i-\theta_i)^2+(n_i-1)S_i^2+2b_i\gamma}{2\sigma_i^2}-\frac{(\theta_i-\boldsymbol{z}_i^t\boldsymbol{\beta})^2}{2\tau^2}\right\}. \end{aligned} \tag{6.4}$$

We state our main result, which provides a sufficient condition for the propriety of the posterior distribution. To this end, we define $\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m)$.

Theorem 6.1. (a) *The marginal posterior density* $\pi(\boldsymbol{\beta}, \tau^2, \gamma|D)$ *is proper if* $m > p + 2$, $n_i > 1$ *and* $\mathrm{rank}(\boldsymbol{Z}) = p$.
(b) *The model parameters* $\boldsymbol{\beta}, \tau^2$ *and* γ *have finite posterior variances if* $m > p + 6$, $n_i > 1$ *and* $\mathrm{rank}(\boldsymbol{Z}) = p$.

Part (a) of Theorem 6.1 says that the marginal posterior densities of the small area means are proper and part (b) establishes a sufficient condition for obtaining finite measures of uncertainty for the model parameters. We note that the sufficient condition given in Theorem 6.1 is the same as the condition given in Arima et al. (2015) except for $n_i > 1$, where they suggested Bayesian estimators for small area models with measurement errors in covariates. The proof of Theorem 6.1 is deferred to Section 6.5.

Since the posterior distribution in (6.4) cannot be obtained in a closed form, we rely on the Markov chain Monte Carlo technique, in particular the Gibbs sampler, in order to draw samples from the posterior distribution. This requires generating samples from the full conditional distributions of each of $(\theta_1, \ldots, \theta_m, \sigma_1^2, \ldots, \sigma_m^2, \boldsymbol{\beta}, \tau^2)$ given the remaining parameters

and the data D. From the expression given in (6.4), the full conditional distributions are given by

$$\begin{aligned}
&\theta_i|\boldsymbol{\beta},\tau^2,\boldsymbol{\sigma}^2,\boldsymbol{\phi}_{(-i)},\gamma,D \sim N\left(\frac{\tau^2 X_i+\sigma_i^2\boldsymbol{z}_i^t\boldsymbol{\beta}}{\tau^2+\sigma_i^2},\frac{\tau^2\sigma_i^2}{\tau^2+\sigma_i^2}\right), \quad i=1,\ldots,m\\
&\sigma_i^2|\boldsymbol{\beta},\tau^2,\boldsymbol{\sigma}^2_{(-i)},\boldsymbol{\phi},\gamma,D \sim IG\left(\frac{n_i}{2}+a_i,\frac{1}{2}(X_i-\theta_i)^2+\frac{1}{2}(n_i-1)S_i^2+b_i\gamma\right), \quad i=1,\ldots,m\\
&\boldsymbol{\beta}|\tau^2,\boldsymbol{\sigma}^2,\boldsymbol{\phi},\gamma,D \sim N_p\left((\boldsymbol{Z}^t\boldsymbol{Z})^{-1}\boldsymbol{Z}^t\boldsymbol{\phi},\tau^2(\boldsymbol{Z}^t\boldsymbol{Z})^{-1}\right),\\
&\tau^2|\boldsymbol{\beta},\boldsymbol{\sigma}^2,\boldsymbol{\phi},\gamma,D \sim IG\left(\frac{m}{2}-1,\frac{1}{2}(\boldsymbol{\phi}-\boldsymbol{Z}\boldsymbol{\beta})^t(\boldsymbol{\phi}-\boldsymbol{Z}\boldsymbol{\beta})\right),\\
&\gamma|\boldsymbol{\beta},\tau^2,\boldsymbol{\sigma}^2,\boldsymbol{\phi},D \sim \Gamma\left(\sum_{i=1}^m a_i+1,\sum_{i=1}^m\frac{b_i}{\sigma_i^2}\right),
\end{aligned} \tag{6.5}$$

where $\boldsymbol{\sigma}^2=(\sigma_1^2,\ldots,\sigma_m^2)^t$, $\boldsymbol{\phi}=(\theta_1,\ldots,\theta_m)^t$, and the suffix $(-i)$ denotes the vector without the i-th component. Fortunately, the full conditional distributions for every parameter are familiar distributions allowing us to easily implement the Gibbs sampling.

In closing of this section, we give a suggestion for the choice of a_i and b_i. For fixed value of γ, it is noted that

$$\mathrm{Var}(X_i)=\mathrm{E}[\mathrm{Var}(X_i|\theta_i)]+\mathrm{Var}(\mathrm{E}[X_i|\theta_i])=\mathrm{E}[\sigma_i^2]=\frac{b_i}{a_i-1}\gamma.$$

Since X_i is the sample mean, it is natural to consider $\mathrm{Var}(X_i)=O(n_i^{-1})$. On the other hand, the full conditional expectation of σ_i^2 is obtained from (6.5) as

$$\begin{aligned}
E[\sigma_i^2|X_i,\theta_i,S_i^2] &= \frac{(X_i-\theta_i)^2/2+(n_i-1)S_i^2/2+b_i\gamma}{n_i/2+a_i-1}\\
&= \frac{n_i/2}{n_i/2+a_i-1}\widetilde{\sigma}_i^2(X_i,S_i^2)+\frac{a_i-1}{n_i/2+a_i-1}\cdot\frac{b_i}{a_i-1}\gamma
\end{aligned}$$

where

$$\widetilde{\sigma}_i^2(X_i,S_i^2)=\frac{1}{n_i}\left\{(X_i-\theta_i)^2+(n_i-1)S_i^2\right\}.$$

Thus the full conditional expectation of σ_i^2 is the weighted mean of $\widetilde{\sigma}_i^2(X_i,S_i^2)$ and the prior mean $b_i\gamma/(a_i-1)$, and the weight for the prior mean is determined by a_i. It is natural that the posterior full conditional expectation approaches to S_i for large n_i. Thus it is reasonable to choose a_i as $a_i=O(1)$ for n_i. These observations show that the order of a_i and b_i should be $a_i=O(1)$ and $b_i=(n_i^{-1})$. Hence, we suggest to use $a_i=2$ and $b_i=n_i^{-1}$ as the one reasonable choice. In the simulation and empirical studies given in the subsequent section, we use these values for a_i and b_i. In empirical study, we investigate the influence of choices of a_i and b_i.

6.2.2 Alternative formulation of heteroscedastic variances

We next suggest the alternative formulation of heteroscedastic variances σ_i^2 in each area. Remember that we assume that $\sigma_i^2 \sim IG(a_i,b_i\gamma)$ for specified a_i and b_i in the previous

subsection. However, in case that we can accommodate the covariate information in the variance modeling, more sophisticated modeling can be developed. Let $\boldsymbol{w}_i$ be a vector of q covariates in the i-th area and $\boldsymbol{\eta}$ is a q-dimensional vector of unknown coefficients, and we propose the structure $\sigma_i^2 \sim IG\left(a_i, b_i\gamma\exp(\boldsymbol{w}_i^t\boldsymbol{\eta})\right)$ with typical choice $a_i = 2$ and $b_i = 1/n_i$. Let $\boldsymbol{w}_i = (w_{i1}, \ldots, w_{iq})^t$ and $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_q)^t$, then we cannot assign $w_{i1} = 1$ for $i = 1, \ldots, m$ since we cannot identify γ and η_1 in this case. To develop a Bayesian inference, we again use the uniform prior distribution for all parameters $\boldsymbol{\beta}, \tau^2, \gamma$ and $\boldsymbol{\eta}$, namely $\pi(\boldsymbol{\beta}, \tau^2, \gamma, \boldsymbol{\eta}) \propto 1$, to keep objectivity of inferences. Therefore, the covariate dependent version of (6.3) is given by

$$\begin{aligned} &X_i|\theta_i, \sigma_i^2 \sim N(\theta_i, \sigma_i^2), \quad \theta_i|\boldsymbol{\beta}, \tau^2 \sim N(\boldsymbol{z}_i^t\boldsymbol{\beta}, \tau^2), \\ &S_i^2|\sigma_i^2 \sim \Gamma\left(\frac{n_i-1}{2}, \frac{n_i-1}{2\sigma_i^2}\right), \quad \sigma_i^2 \sim IG\left(a_i, b_i\gamma\exp(\boldsymbol{w}_i^t\boldsymbol{\eta})\right) \\ &\pi(\boldsymbol{\beta}, \tau^2, \gamma, \boldsymbol{\eta}) \propto 1, \end{aligned} \tag{6.6}$$

Then, the joint posterior distribution (6.4) is changed as

$$\begin{aligned} \pi(\theta_1, \ldots, \theta_m, \sigma_1^2, \ldots, \sigma_m^2, \boldsymbol{\beta}, \tau^2, \gamma, \boldsymbol{\eta}|D) &\propto (\tau^2)^{-m/2}\prod_{i=1}^m \gamma^{a_i}\exp(a_i\boldsymbol{w}_i^t\boldsymbol{\eta})(\sigma_i^2)^{-n_i/2-a_i-1} \\ &\times \exp\left\{-\frac{(X_i-\theta_i)^2 + (n_i-1)S_i^2 + 2b_i\gamma\exp(\boldsymbol{w}_i^t\boldsymbol{\eta})}{2\sigma_i^2} - \frac{(\theta_i - \boldsymbol{z}_i^t\boldsymbol{\beta})^2}{2\tau^2}\right\}. \end{aligned} \tag{6.7}$$

We state our second main result, which provides a sufficient condition for the propriety of the posterior distribution given in (6.7). To this end, we define

$$t_k = \mathrm{sgn}\left(\sum_{i=1}^m a_i w_{ik}\right)\mathrm{sgn}\left(\sum_{i=1}^m n_i w_{ik}\right), \quad k = 1, \ldots, q,$$

where $\mathrm{sgn}(x)$ for the real number x denotes the sign of x.

Theorem 6.2. (a) *The marginal posterior density $\pi(\boldsymbol{\beta}, \tau^2, \gamma, \boldsymbol{\eta}|D)$ is proper if $m > p + 2$, $n_i > 1$, $\mathrm{rank}(\boldsymbol{Z}) = p$, and $t_k = 1$ for $k = 1, \ldots, q$.*
(b) *The model parameters $\boldsymbol{\beta}, \tau^2, \gamma$ and $\boldsymbol{\eta}$ have finite posterior variances if $m > p+6$, $n_i > 1$, $\mathrm{rank}(\boldsymbol{Z}) = p$, and $t_k = 1$ for $k = 1, \ldots, q$.*

The last new condition $t_k = 1$ for $k = 1, \ldots, q$ given in both (a) and (b) means that the two values $\sum_{i=1}^m a_i w_{ik}$ and $\sum_{i=1}^m n_i w_{ik}$ have the same signs for $k = 1, \ldots, q$, while other conditions are the same as in Theorem 6.1. Note that the simple sufficient condition for the last condition is w_{ik}, $i = 1, \ldots, m$ have the same signs since a_i and n_i are positive.

To sample from the joint posterior distribution (6.7), we can again use the Gibbs sampling method. Note that the full conditional distributions of θ_i's, $\boldsymbol{\beta}$ and τ^2 are the same as (6.5), and these of σ_i^2 and γ are obtained by replacing b_i with $\exp(\boldsymbol{w}_i^t\boldsymbol{\eta})$. The full conditional distribution of $\boldsymbol{\eta}$ is proportional to

$$\pi(\boldsymbol{\eta}|\boldsymbol{\sigma}^2, \gamma, D) = \prod_{i=1}^m \exp(a_i\boldsymbol{w}_i^t\boldsymbol{\eta})\exp\left\{-\frac{b_i\gamma\exp(\boldsymbol{w}_i^t\boldsymbol{\eta})}{\sigma_i^2}\right\},$$

which is not a familiar form. To sample from this full conditional distribution, we use the random-walk Metropolis-Hastings (MH) algorithm. Let $\boldsymbol{\eta}_0$ be the current value and we generate the proposal $\boldsymbol{\eta}^*$ from $N_q(\boldsymbol{\eta}_0, c\boldsymbol{I}_q)$ for specified $c > 0$. Then we accept the proposal $\boldsymbol{\eta}^*$ with probability $\min\{1, p(\boldsymbol{\eta}_0, \boldsymbol{\eta}^*)\}$, where

$$p(\boldsymbol{\eta}_0, \boldsymbol{\eta}^*) = \prod_{i=1}^{m} \exp\{a_i \boldsymbol{w}_i^t(\boldsymbol{\eta}^* - \boldsymbol{\eta}_0)\} \exp\left(\frac{-b_i\gamma[\exp(\boldsymbol{w}_i^t\boldsymbol{\eta}^*) - \exp(\boldsymbol{w}_i^t\boldsymbol{\eta}_0)]}{\sigma_i^2}\right).$$

6.3 Simulation studies

In this section, we compare the accuracy of the hierarchical Bayes estimator based on the proposed full Bayesian model with the empirical Bayes estimator given by Maiti et al. (2014) and the hierarchical model suggested in You and Chapman (2006) through simulation experiments. We first generate observations for each small area from

$$X_{ij} = \beta_0 + \beta_1 z_i + u_i + e_{ij}, \quad j = 1, \ldots, n_i,\ i = 1, \ldots, m,$$

where $u_i \sim N(0, \tau^2)$ and $e_{ij} \sim N(0, n_i\sigma_i^2)$. Then the random effects model for the small area mean is

$$X_i = \beta_0 + \beta_1 z_i + u_i + e_i, \quad i = 1, \ldots, m,$$

where $X_i = \overline{X}_i = n_i^{-1}\sum_{j=1}^{n_i} X_{ij}$ and $e_i = n_i^{-1}\sum_{j=1}^{n_i} e_{ij}$. Therefore, $X_i|\theta_i \sim N(\theta_i, \sigma_i^2)$, where $\theta_i = \beta_0 + \beta_1 z_i + u_i$, that is $\theta_i \sim N(\beta_0 + \beta_1 z_i,)$, and $e_i \sim N(0, \sigma_i^2)$. The parameter of interest is the mean θ_i in the i-th small area. The direct estimator of σ_i^2 we used in simulation runs is

$$S_i^2 = \frac{1}{n_i(n_i - 1)} \sum_{j=1}^{n_i} (X_{ij} - \overline{X}_i)^2,$$

noting that $S_i^2|\sigma_i^2 \sim \Gamma((n_i - 1)/2, (n_i - 1)/2\sigma_i^2)$. We generate covariate z_i from the uniform distribution on $(2, 8)$, and set the true parameter values $\beta_0 = 0.5, \beta_1 = 0.8$ and $\tau^2 = 1$. We consider the case $m = 30$ and $n_i = 7$ for all areas. For the true values of σ_i^2, we consider two cases: (i) $\sigma_i^2 \sim IG(10, 5\exp(0.3z_i))$ and (ii) $\sigma_i^2 \sim U(0.5, 5)$.

For simulated data, we apply four methods to get the estimator of the small area mean θ_i and variance σ_i^2. Two of four are the proposed Bayesian models (6.3) and (6.6) referred as STK1 and STK2, respectively. In applying these models, we put $a_i = 2$ and $b_i = 1/n_i$ as discussed in the end of Section 6.2, and we use $c = (0.2)^2$ in each MH step in STK2. The third method is the hierarchical Bayesian method given by You and Chapman (2006) referred to as YC, where we assign the uniform prior for σ_i^2, namely $\pi(\sigma_i^2) \propto 1$. For posterior sampling in YC method, we replace the full conditional for σ_i^2 in (6.5) with

$$\sigma_i^2|\boldsymbol{\beta}, \tau^2, \boldsymbol{\sigma}_{(-i)}^2, \boldsymbol{\phi}, D \sim IG\left(\frac{n_i}{2}, \frac{1}{2}(X_i - \theta_i)^2 + \frac{1}{2}(n_i - 1)S_i^2\right), \quad i = 1, \ldots, m,$$

and the propriety of the posterior distribution can be easily established from small modification of the proof of Theorem 6.1. The fourth method is the empirical Bayes method given in Maiti et al. (2014) referred to as MRS. In the three full Bayesian model, we calculate the

estimators $\widehat{\theta}_i$ and $\widehat{\sigma}_i^2$ as the mean of 5000 posterior samples after 1000 iteration. For all four estimator, we calculate the mean squared errors and the absolute biases defined as

$$\text{MSE} = \frac{1}{mR}\sum_{i=1}^{m}\sum_{r=1}^{R}(\widehat{\theta}_i^{(r)} - \theta_i^{(r)})^2, \quad \text{Bias} = \frac{1}{mR}\sum_{i=1}^{m}\left|\sum_{r=1}^{R}(\widehat{\theta}_i^{(r)} - \theta_i^{(r)})\right|,$$

based on $R = 2000$ simulation runs, where $\widehat{\theta}_i^{(r)}$ and $\theta_i^{(r)}$ are the estimated and true value in the i-th area in the r-th iteration. Moreover, for the three Bayesian models STK1, STK2 and YC, we compute the credible intervals of θ_i with probability 0.95 and 0.99, and calculated the coverage probability $(mR)^{-1}\sum_{i=1}^{m}\sum_{r=1}^{R} I(\theta_i \in \widehat{\text{CI}_{i(r)}})$, where $\widehat{\text{CI}_{i(r)}}$ denotes the credible interval for θ_i in the r-th run. The simulation results are presented in Table 7.1. For point estimation of θ_i, the MSEs of θ_i in MRS is reasonable values, but the bias of MRS is larger compared to other three Bayesian models. Among the full Bayesian model, it is observed that STK1 and STK2 attain minimum values of MSE in the case (ii) and (i), respectively. The preference of YC is worst among the four models since YC does not consider the shrinkage estimation of σ_i^2 in spite of small sample sizes ($n_i = 7$). We also noted that the MSEs of σ_i^2 are largest in MRS in both cases, which may comes from instability of estimation of α and γ in (6.2). Concerned with the Bayesian credible intervals, it is revealed that the suggested two methods STK1 and STK2 almost attain the nominal levels, but YC provides smaller coverage provabilities than the nominal levels. This is clear that this phenomena comes from the instability of variance estimation in the YC method. Therefore, the suggested procedure reasonably works in terms of MSE and bias of both θ_i and σ_i^2, and can provide an accurate credible interval compared to the YC method.

Table 6.1: Simulation Result.

		Mean (θ_i)		Variance (σ_i^2)		CP	
		MSE	Bias	MSE	Bias	95%	99%
(i)	STK1	1.120	0.036	2.325	0.411	95.6	99.3
	STK2	1.102	0.035	2.087	0.272	95.3	99.2
	YC	1.275	0.038	3.894	0.120	93.2	97.6
	MRS	1.149	0.410	4.442	0.451	—	—
(ii)	STK1	1.043	0.040	1.144	0.041	95.2	99.2
	STK2	1.053	0.041	1.845	0.278	95.5	99.4
	YC	1.185	0.044	2.630	0.099	93.0	97.9
	MRS	1.001	0.273	2.849	0.320	—	—

6.4 Real Data Analysis

6.4.1 Survey data

We apply the suggested procedures to the data in the Survey of Family Income and Expenditure (SFIE) in Japan. In this study, we use the data of the spending item 'Education' (scaled

by 1000) in the survey in November 2011. The average spending at each capital city of 47 prefectures in Japan is denoted by X_i for $i = 1, \ldots, 47$. Although the average spendings in SFIE are reported every month, the sample sizes n_i's are around 100 for most prefectures, and data of the item 'Education' have high variability. On the other hand, we have data in the National Survey of Family Income and Expenditure (NSFIE) for 47 prefectures. Since NSFIE is based on much larger sample than SFIE, the average spendings in NSFIE are more reliable, but this survey has been implemented every five years. In this study, we use the data of the item 'Education' of NSFIE in 2009 as a covariate, which is denoted by z_i for $i = 1, \ldots, 47$. Then the two stage model for X_i is described as

$$X_i|\theta_i, \sigma_i^2 \sim N(\theta_i, \sigma_i^2), \quad \theta_i|\beta_0, \beta_1, \tau^2 \sim N(\beta_0 + \beta_1 z_i, \tau^2), \quad i = 1, \ldots, 47.$$

As the direct estimates of σ_i^2, we calculate S_i^2 from the data of the spending 'Education' at the same city every November in the past ten years. Then the model for S_i^2 is given by

$$S_i^2|\sigma_i^2 \sim \Gamma\left(\frac{n_i - 1}{2}, \frac{n_i - 1}{2\sigma_i^2}\right), \quad i = 1, \ldots, 47.$$

and the priors for σ_i^2 are given by

$$\text{(STK1) } \sigma_i^2 \sim IG(a_i, b_i\gamma), \quad \text{(STK2) } \sigma_i^2 \sim IG(a_i, b_i\gamma \exp(\eta z_i)), \quad \text{(YC) } \pi(\sigma_i^2) \propto 1.$$

Remember that the uniform prior for σ_i^2 in YC model leads to the non-shrinkage posterior estimator of σ_i^2, while the proper prior for σ_i^2 in STK1 and STK2 leads to the shrinkage estimator of σ_i^2 toward the prior mean.

It is easy to confirm that the sufficient conditions in Theorems 6.1 and 6.2 are satisfied in this case since the covariate z_i is positive for all areas. Now, we apply the three models to the survey data with

$$a_i = 2 \quad \text{and} \quad b_i = 1/n_i \quad \text{in STK1 and STK2.}$$

Moreover, to investigate sensitivity of the choices of a_i and b_i, we consider the following two additional choices:

$$\text{(s1)} \quad a_i = 3, \ b_i = 1/n_i, \qquad \text{(s2)} \quad a_i = 2, \ b_i = 1, \tag{6.8}$$

where the prior mean of σ_i^2 is $\gamma/(2n_i)$ and γ in (s1) and (s2), respectively. We use $c = 1$ for MH step in STK2. We first calculate the point estimates of model parameters as the means of 95000 posterior samples by Gibbs sampling after 5000 iteration. The results are given in Table 6.2. The estimated values of β_0, β_1 and τ^2 are similar for all models. For model comparison of these models, we calculated the Deviance Information Criterion (DIC) of Spiegelhalter et al. (2002) given by $\text{DIC} = 2\overline{D(\phi)} - D(\overline{\phi})$, where $\boldsymbol{\phi}$ is the unknown model parameters, $D(\phi)$ is (-2) times log-marginal likelihood function, and $\overline{D(\phi)}$ and $\overline{\phi}$ denote that posterior means of $D(\phi)$ and ϕ, respectively. Note that $\phi = \{\boldsymbol{\beta}, \tau^2, \gamma\}$ for STK1, $\phi = \{\boldsymbol{\beta}, \tau^2, \gamma, \boldsymbol{\eta}\}$ for STK2, and $\phi = \{\boldsymbol{\beta}, \tau^2, \sigma_1^2, \ldots, \sigma_m^2\}$ for YC. The resulting values of DIC and $\overline{D(\phi)}$ are reported in Table 6.2, and it is observed that YC is the most suitable model for this data set in terms of DIC. This may come from the fact that the sample size n_i in each area is around 100. Thus the direct estimates of sampling variances are relatively accurate in this case, so that it does

not require shrinkage estimation for variances. Comparing STK1, STK1-(s1) and STK1-(s2), γ seems sensitive to the choice of a_i and b_i, since the prior means are different for each choice, but the recommended choice attains the smaller value of DIC. The same thing can be observed in STK2, STK2-(s1) and STK2-(s2). However, the posterior mean of θ_i and σ_i^2 are nearly the same among the three choices.

In the closing of this study, we compute the posterior estimates of σ_i^2's and θ_i^2's obtained from three models, STK1, STK2 and YC. In Figure 6.1, we provide the scatter plots of direct and posterior estimates of σ_i^2's and θ_i's for selected 15 areas. From the left panel of Figure 6.1, the posterior estimates of σ_i^2 are almost the same for each model in the area with small direct estimates. On the other hand, in areas with large direct estimates of σ_i^2, the posterior estimates in YC and those of STK1 or STK2 are different since STK1 and STK2 produce shrinkage estimators for σ_i^2, but the difference is still small. For the scatter plot for θ_i given in the right panel of Figure 6.1, it is observed that the resulting posterior estimates from three models are similar. Thus, the suggested procedures STK1 and STK2 provide almost the same estimates of θ_i, parameter of interest, as the YC method while the DIC values of STK1 and STK2 are larger than YC. That is, both STK1 and STK2 work as well as YC in the case that there are no need to shrink direct estimates of variances.

Table 6.2: Posterior Points Estimates and Standard Errors (Parenthesis) of Model Parameters, and DICs in Survey Data.

	β_0	β_1	τ^2	γ	η	DIC
STK1	0.893	0.696	10.5	2.42×10^3	—	700.4
	(2.74)	(0.206)	(5.15)	(2.57×10^2)	—	
STK1-(s1)	0.864	0.698	10.5	3.71×10^3	—	717.4
	(2.76)	(0.207)	(5.15)	(3.29×10^2)	—	
STK1-(s2)	0.918	0.694	10.4	23.6	—	705.0
	(2.77)	(0.207)	(5.12)	(2.52)	—	
STK2	0.868	0.697	10.4	1.22×10^3	5.47×10^{-2}	700.3
	(2.78)	(0.209)	(5.17)	(3.08×10^2)	(1.74×10^{-2})	
STK2-(s1)	0.878	0.697	10.4	2.69×10^3	2.22×10^{-2}	745.3
	(2.77)	(0.208)	(5.18)	(4.07×10^2)	(1.08×10^{-2})	
STK2-(s2)	0.831	0.700	10.6	30.6	-1.68×10^{-2}	6651.8
	(2.77)	(0.208)	(5.19)	(10.7)	(2.63×10^{-2})	
YC	0.913	0.698	11.0	—	—	**558.3**
	(2.74)	(0.206)	(5.25)	—	—	

6.4.2 Corn data

We next illustrate our methods based on the widely studied example which was first analyzed by Battese et al. (1988). The dataset is on corn and soybean productions in 12 Iowa counties, and we here focus on corn data. Since the sample size of the original data is ranging from 1 to 5, we cannot use the proposed model which requires $n_i > 1$ for the posterior propriety

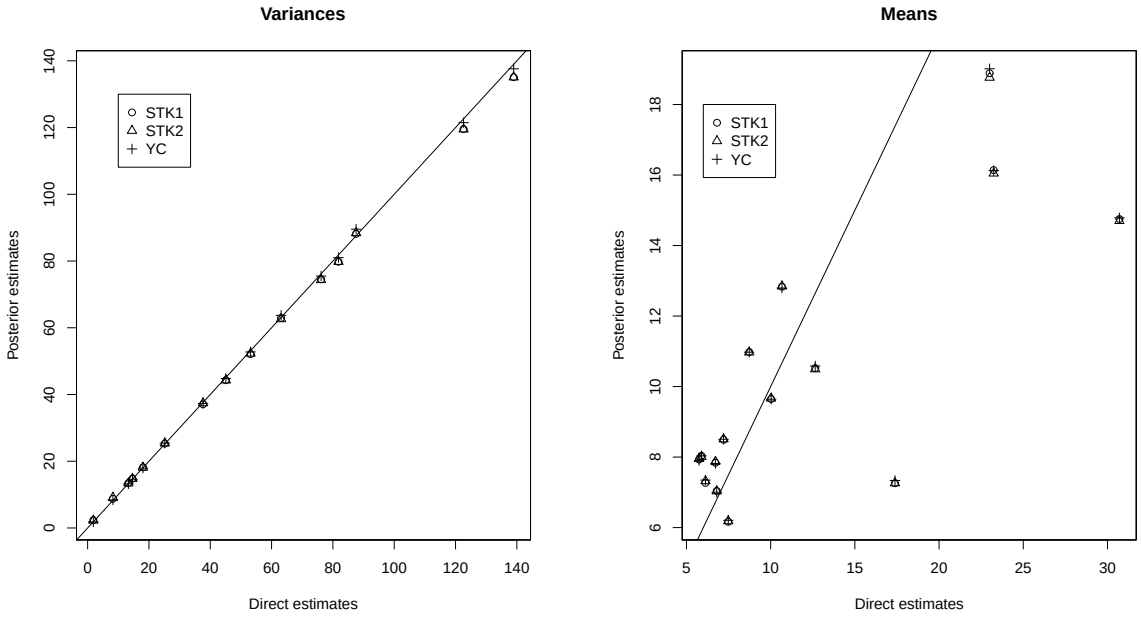

Figure 6.1: Scatter Plots of Direct and Posterior Estimates of σ_i^2's (Left) and θ_i's (Right) for Selected 15 Areas in Survey Data.

as given in Theorem 6.1. Thus, we use the modified data given in the table 6 in Dass et al. (2012). The dataset consists of $m = 8$ areas with sample sizes in each area ranging from 3 to 5, and the survey data of corn (X_i) and the satellite data of both corn (z_{1i}) and soybeans (z_{2i}) as the covariates are observed in each area, where X_i, z_{1i}, z_{2i} are scaled by 100. Note that the sample sizes n_i in each area is much smaller than that in the previous study. Similarly to the previous study, we apply the three models STK1, STK2 with $a_i = 2$ and $b_i = 1/n_i$ and YC. The two stage model for X_i is given by

$$X_i|\theta_i, \sigma_i^2 \sim N(\theta_i, \sigma_i^2), \quad \theta_i|\beta_0, \beta_1, \beta_2, \tau^2 \sim N(\beta_0 + \beta_1 z_{1i} + \beta_2 z_{2i}, \tau^2), \quad i = 1, \ldots, 8.$$

For a covariate for variance modeling in STK2, we use only z_{1i}, namely $\sigma^2 \sim IG(a_i, b_i\gamma \exp(\eta z_{1i}))$, since the DIC values of other models with use of only z_{2i} and both z_{1i} and z_{2i} are larger than this model. Since the covariate z_{1i} is positive for all areas, the sufficient conditions in Theorem 6.1 and 6.2 are satisfied in this case. We use $c = (0.2)^2$ in each MH step in STK2. We again consider two additional choices of a_i and b_i in (6.8). Then, based on 95000 posterior samples after 5000 iteration, we calculate the point estimates of model parameters as the posterior sample means and we provide the resulting values in Table 6.3 as well as DIC values. The posterior estimates of regression coefficients β_0, β_1 and β_2 are similar for all models, but γ and η are different depending on the choices of a_i and b_i. It is also revealed that STK2 is the most preferable model for this data set from DIC values. Among the three choices of a_i and b_i, the recommended choice seems the best in terms of DIC, but the posterior mean of θ_i and σ_i^2 are almost the same among the three choices. In this case, it is interesting to point out that both STK1 and STK2 are more preferable than YC in terms of DIC values. This is

because the accuracy of the direct estimates of variances with small sample sizes (from 3 to 5) is suspicious and the shrinkage estimation for σ_i^2 is needed in this case.

In the left panel of Figure 6.2, we show the scatter plots of direct and posterior estimates of σ_i^2 obtained from three models, STK1, STK2, and YC. The result shows that the posterior estimates of σ_i^2 of YC (using uniform prior on σ_i^2) are considerably different from those of STK1 or STK2, while STK1 and STK2 produce the similar posterior estimated values. It is also observed that the posterior estimator of σ_i^2 of STK1 and STK2 shrink the direct estimator of σ_i^2 toward some prior mean, but that of YC does not. In the right panel of Figure 6.2, we show the 95% credible intervals for θ_i from each model. It is clear that STK1 and STK2 produce similar credible intervals and YC produces shorter credible intervals than two methods since the length of credible intervals are affected by the posterior estimates of σ_i^2. In particular, the credible interval of YC in area 1 is much shorter than that of STK1 and STK2, but the interval of YC is not reliable because of instability of variance estimation in the YC method. Then we may misinterpret the accuracy of the resulting estimator of θ_i when we use YC in this case. This phenomena is consistent to the simulation results in Table 7.1, where the credible interval in YC has smaller coverage probability than the true nominal level. Thus the shrinking variances is the crucial strategy when n_i is small like this data set.

Table 6.3: Posterior Points Estimates and Standard Errors (Parenthesis) of Model Parameters, and DICs in Corn Data.

	β_0	β_1	β_2	τ^2	γ	η	DIC
STK1	-1.59	0.679	0.379	0.278	0.559	—	-14.45
	(9.47)	(1.97)	(1.88)	(1.25)	(0.252)	—	
STK1-(s1)	-1.42	0.643	0.347	0.279	0.884	—	-14.23
	(9.68)	(2.02)	(1.89)	(1.69)	(0.367)	—	
STK1-(s2)	-1.73	0.726	0.385	0.341	0.144	—	-11.76
	(9.67)	(2.03)	(1.90)	(2.63)	(0.0655)	—	
STK2	-1.76	0.720	0.402	0.367	8.67	-0.939	**-20.39**
	(11.0)	(2.38)	(2.03)	(7.21)	(4.77)	(0.154)	
STK2-(s1)	-1.57	0.686	0.358	0.256	14.5	-0.961	-10.02
	(9.06)	(1.90)	(1.77)	(0.821)	(7.05)	(0.118)	
STK2-(s2)	-1.74	0.729	0.384	0.283	7.31	-1.27	-3.10
	(9.53)	(1.99)	(1.88)	(1.20)	(5.56)	(0.299)	
YC	-1.805	0.754	0.375	0.303	—	—	-7.33
	(9.57)	(1.99)	(1.88)	(1.11)	—	—	

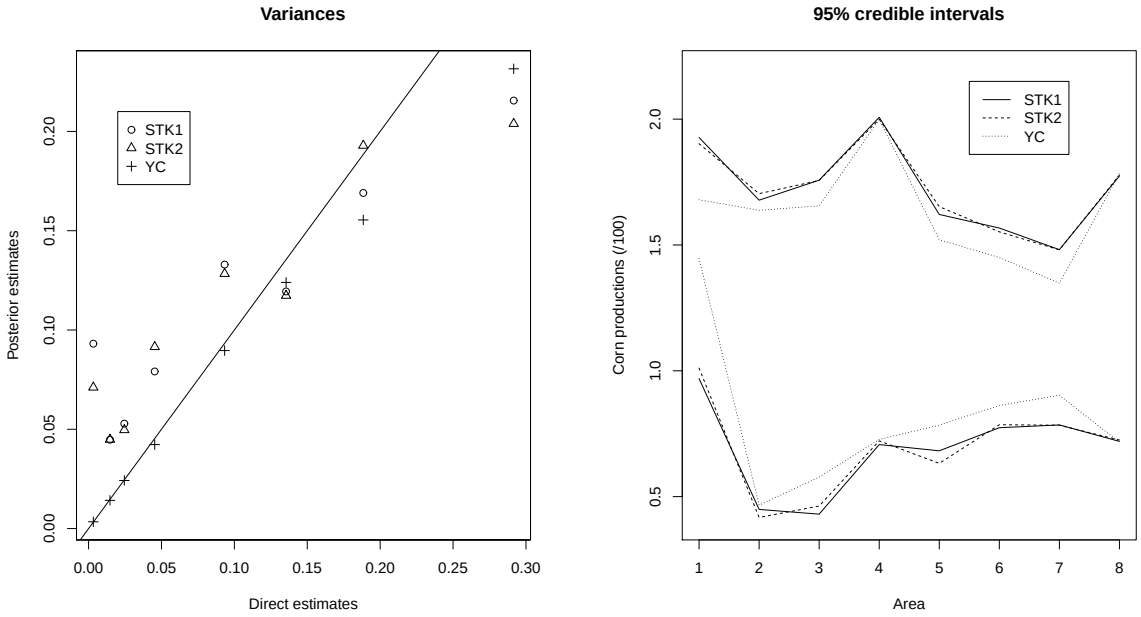

Figure 6.2: Scatter Plots of Direct and Posterior Estimates of σ_i^2's (Left) and 95% Credible Intervals of θ_i's (Right) in Corn Data.

6.5 Proofs

6.5.1 Proof of Theorem 6.1.

We first prove part (a). Let $\mathbb{R}_+ = \{x \in \mathbb{R} |\ x > 0\}$ be the set of positive numbers. In what follows, capital C, with and without suffix, means a generic constant. It is sufficient to prove that

$$\int_{\mathbb{R}^m \times \mathbb{R}_+^2} \pi(\boldsymbol{\beta}, \tau^2, \gamma | D) d\boldsymbol{\beta} d\tau^2 d\gamma < \infty.$$

Let $\boldsymbol{\phi} = (\theta_1, \ldots, \theta_m)^t$, $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_m^2)^t$. Then we need to prove that

$$\int_{\mathbb{R}^m \times \boldsymbol{R}_+^m \times \boldsymbol{R}^p \times \mathbb{R}_+^2} \pi(\theta_1, \ldots, \theta_m, \sigma_1^2, \ldots, \sigma_m^2, \boldsymbol{\beta}, \tau^2, \gamma | D) d\boldsymbol{\phi} d\boldsymbol{\sigma}^2 d\boldsymbol{\beta} d\tau^2 d\gamma < \infty,$$

where

$$\pi(\theta_1, \ldots, \theta_m, \sigma_1^2, \ldots, \sigma_m^2, \boldsymbol{\beta}, \tau^2, \gamma | D)$$
$$\propto (\tau^2)^{-m/2} \prod_{i=1}^m \gamma^{a_i} (\sigma_i^2)^{-n_i/2 - a_i - 1} \exp\left(-\frac{(X_i - \theta_i)^2 + (n_i - 1) S_i^2 + 2 b_i \gamma}{2\sigma_i^2} - \frac{(\theta_i - \boldsymbol{z}_i^t \boldsymbol{\beta})^2}{2\tau^2} \right).$$

From expression (6.4), we first integrate with respect to $\sigma_1^2, \ldots, \sigma_m^2$ to get

$$\pi(\boldsymbol{\phi}, \boldsymbol{\beta}, \tau^2, \gamma | D) \propto (\tau^2)^{-m/2} \exp\left(-\frac{(\boldsymbol{\phi} - \boldsymbol{Z}^t \boldsymbol{\beta})^t (\boldsymbol{\phi} - \boldsymbol{Z}^t \boldsymbol{\beta})}{2\tau^2} \right) \prod_{i=1}^m \gamma^{a_i} \psi_i(\theta_i - X_i, \gamma)^{-(n_i/2 + a_i)},$$

where $\psi_i(\theta_i - X_i, \gamma) = (X_i - \theta_i)^2 + (n_i - 1)S_i^2 + 2b_i\gamma$. Noting that

$$\int_{\mathbb{R}^p} \exp\left(-\frac{(\boldsymbol{\theta} - \boldsymbol{Z}\boldsymbol{\beta})^t(\boldsymbol{\theta} - \boldsymbol{Z}^t\boldsymbol{\beta})}{2\tau^2}\right) d\boldsymbol{\beta} = (\tau^2)^{p/2}|\boldsymbol{Z}^t\boldsymbol{Z}|^{-1/2} \exp\left\{-\frac{1}{2\tau^2}\boldsymbol{\theta}^t(\boldsymbol{I}_m - \boldsymbol{Z}(\boldsymbol{Z}^t\boldsymbol{Z})^{-1}\boldsymbol{Z}^t)\boldsymbol{\theta}\right\},$$

we obtain

$$\pi(\boldsymbol{\theta}, \tau^2, \gamma|D) \propto (\tau^2)^{-(m-p-2)/2-1} \exp\left\{-\frac{1}{2\tau^2}\boldsymbol{\theta}^t\boldsymbol{A}\boldsymbol{\theta}\right\} \prod_{i=1}^m \gamma^{a_i}\psi_i(\theta_i - X_i, \gamma)^{-(n_i/2+a_i)}, \quad (6.9)$$

for $\boldsymbol{A} = \boldsymbol{I}_m - \boldsymbol{Z}(\boldsymbol{Z}^t\boldsymbol{Z})^{-1}\boldsymbol{Z}^t$. When $m - p - 2 > 0$ i.e. $m > p + 2$, we can integrate (6.9) with respect to τ^2 to get

$$\pi(\boldsymbol{\theta}, \gamma|D) \propto \left(\boldsymbol{\theta}^t\boldsymbol{A}\boldsymbol{\theta}\right)^{-(m-p-2)/2} \prod_{i=1}^m \gamma^{a_i}\psi_i(\theta_i - X_i, \gamma)^{-(n_i/2+a_i)}. \quad (6.10)$$

Define $\Omega = \{\boldsymbol{\theta}|\boldsymbol{\theta}^t\boldsymbol{A}\boldsymbol{\theta} \le 1\} \subset \mathbb{R}^m$. It holds that

$$\int_{\mathbb{R}^m\times\mathbb{R}_+} \pi(\boldsymbol{\theta}, \gamma|D)d\boldsymbol{\theta}d\gamma = \int_{\Omega\times\mathbb{R}_+} \pi(\boldsymbol{\theta}, \gamma|D)d\boldsymbol{\theta}d\gamma + \int_{\Omega^c\times\mathbb{R}_+} \pi(\boldsymbol{\theta}, \gamma|D)d\boldsymbol{\theta}d\gamma.$$

The second term can be evaluated as

$$\begin{aligned}
\int_{\Omega^c\times\mathbb{R}_+} \pi(\boldsymbol{\theta}, \gamma|D)d\boldsymbol{\theta}d\gamma &\le C \int_{\Omega^c\times\mathbb{R}_+} \prod_{i=1}^m \gamma^{a_i}\psi_i(\theta_i - X_i, \gamma)^{-(n_i/2+a_i)}d\boldsymbol{\theta}d\gamma \\
&\le C \int_{\mathbb{R}^m\times\mathbb{R}_+} \prod_{i=1}^m \gamma^{a_i}\psi_i(\theta_i - X_i, \gamma)^{-(n_i/2+a_i)}d\boldsymbol{\theta}d\gamma \\
&= C \int_0^\infty \prod_{i=1}^m \Big\{ \int_{-\infty}^\infty \gamma^{a_i}\psi_i(\theta_i - X_i, \gamma)^{-(n_i/2+a_i)}d\theta_i \Big\} d\gamma,
\end{aligned}$$

which corresponds to the last formula in (12), and it is finite. For evaluating the first term, we first note that there exists a $(m-p)\times m$ matrix $\boldsymbol{H}_1$ such that $\boldsymbol{A} = \boldsymbol{H}_1^t\boldsymbol{H}_1$ and $\boldsymbol{H}_1\boldsymbol{H}_1^t = \boldsymbol{I}_{m-p}$ since $\boldsymbol{A}$ is an idempotent matrix with $\mathrm{rank}(\boldsymbol{A}) = m-p$. By changing the variable as $\boldsymbol{u}_1 = \boldsymbol{H}_1\boldsymbol{\theta}$ and $\boldsymbol{u}_2 = (u_{m-p+1}, \ldots, u_m)'$ with $u_i = \theta_i$, it follows that

$$\begin{aligned}
\int_{\Omega\times\mathbb{R}_+} \pi(\boldsymbol{\theta}, \gamma|D)d\boldsymbol{\theta}d\gamma \le C' \int_{\boldsymbol{u}_1^t\boldsymbol{u}_1\le 1} (\boldsymbol{u}_1^t\boldsymbol{u}_1)^{-(m-p-2)/2}d\boldsymbol{u}_1 \int_0^\infty \prod_{i=1}^{m-p} \gamma^{a_i}\{(n_i-1)S_i^2 + 2b_i\gamma\}^{-(n_i/2+a_i)} \\
\times \prod_{i=m-p+1}^m \left\{ \int_{-\infty}^\infty \gamma^{a_i}\psi_i(u_i - X_i, \gamma)^{-(n_i/2+a_i)}du_i \right\} d\gamma.
\end{aligned}$$

Moreover, it holds that

$$\int_{\boldsymbol{u}_1^t\boldsymbol{u}_1\le 1} (\boldsymbol{u}_1^t\boldsymbol{u}_1)^{-(m-p-2)/2}d\boldsymbol{u}_1 = C'' \int_0^1 r^{-(m-p-2)}r^{m-p-1}dr < \infty,$$

thereby the similar evaluation shows that $\int_{\Omega\times\mathbb{R}_+} \pi(\boldsymbol{\theta}, \gamma|D)d\boldsymbol{\theta}d\gamma$ is also finite. Thus the proof for part (a) is complete.

For part (b), we show $E(\boldsymbol{\beta}\boldsymbol{\beta}^t|D), E((\tau^2)^2|D)$ and $E(\gamma^2|D)$ are finite. For $E((\tau^2)^2|D)$, we evaluate it in the same manner as in Part (a). Note that

$$(\tau^2)^2\pi(\boldsymbol{\theta},\tau^2,\gamma|D) \propto (\tau^2)^{-(m-p-6)/2-1}\exp\left(-\frac{1}{2\tau^2}\boldsymbol{\theta}^t\boldsymbol{A}\boldsymbol{\theta}\right)\prod_{i=1}^m \gamma^{a_i}\psi_i(\theta_i - X_i,\gamma)^{-(n_i/2+a_i)},$$

so that it follows, when $m-p-6>0$, namely $m>p+6$, that

$$E((\tau^2)^2|D) < C\int_{\mathbb{R}^m\times\mathbb{R}_+}\prod_{i=1}^m \gamma^{a_i}\psi_i(\theta_i - X_i,\gamma)^{-(n_i/2+a_i)}d\boldsymbol{\theta}d\gamma < \infty.$$

For evaluating $E(\boldsymbol{\beta}\boldsymbol{\beta}^t|D)$, note that

$$\begin{aligned}&\int_{\mathbb{R}^p}\boldsymbol{\beta}\boldsymbol{\beta}^t\exp\left(-\frac{(\boldsymbol{\theta}-\boldsymbol{Z}\boldsymbol{\beta})^t(\boldsymbol{\theta}-\boldsymbol{Z}^t\boldsymbol{\beta})}{2\tau^2}\right)d\boldsymbol{\beta}\\ &\quad= (\tau^2)^{p/2}|\boldsymbol{Z}^t\boldsymbol{Z}|^{-1/2}\exp\left(-\frac{1}{2\tau^2}\boldsymbol{\theta}^t\boldsymbol{A}\boldsymbol{\theta}\right)(\boldsymbol{Z}^t\boldsymbol{Z})^{-1}\left\{\tau^2\boldsymbol{I}_m + \boldsymbol{Z}^t\boldsymbol{\theta}\boldsymbol{\theta}^t\boldsymbol{Z}(\boldsymbol{Z}^t\boldsymbol{Z})^{-1}\right\}.\end{aligned}$$

Integrating out it with respect to τ^2, we have

$$\begin{aligned}E(\boldsymbol{\beta}\boldsymbol{\beta}^t|D) \propto &\int_{\mathbb{R}^m\times\mathbb{R}_+}\left(\boldsymbol{\theta}^t\boldsymbol{A}\boldsymbol{\theta}\right)^{-(m-p-4)/2}\prod_{i=1}^m\gamma^{a_i}\psi_i(\theta_i-X_i,\gamma)^{-(n_i/2+a_i)}d\boldsymbol{\theta}d\gamma(\boldsymbol{Z}^t\boldsymbol{Z})^{-1}\\ &+\int_{\mathbb{R}^m\times\mathbb{R}_+}\frac{(\boldsymbol{Z}^t\boldsymbol{Z})^{-1}\boldsymbol{Z}^t\boldsymbol{\theta}\boldsymbol{\theta}^t\boldsymbol{Z}(\boldsymbol{Z}^t\boldsymbol{Z})^{-1}}{(\boldsymbol{\theta}^t\boldsymbol{A}\boldsymbol{\theta})^{-(m-p-2)/2}}\prod_{i=1}^m\frac{\gamma^{a_i}}{\psi_i(\theta_i-X_i,\gamma)^{n_i/2+a_i}}d\boldsymbol{\theta}d\gamma.\end{aligned}$$

The first term can be verified to be finite by using the same arguments used to evaluate (6.10).

For the second term, it is sufficient to show that for $j=1,\ldots,m$,

$$\int_{\mathbb{R}^m\times\mathbb{R}_+}\frac{\theta_j^2}{(\boldsymbol{\theta}^t A\boldsymbol{\theta})^{-(m-p-2)/2}}\prod_{i=1}^m\frac{\gamma^{a_i}}{\psi_i(\theta_i-X_i,\gamma)^{n_i/2+a_i}}d\boldsymbol{\theta}d\gamma<\infty. \tag{6.11}$$

By the same arguments used to evaluate (6.10), the inequality (6.11) is satisfied if

$$\int_0^\infty\Big\{\int_{-\infty}^\infty\frac{\gamma^{a_j}\mu_j^2}{\psi_i(\mu_j,\gamma)^{n_j/2+a_j}}d\mu_j\Big\}\prod_{i\neq j}\Big\{\int_{-\infty}^\infty\frac{\gamma^{a_i}}{\psi_i(\mu_i,\gamma)^{n_i/2+a_i-1}}d\mu_i\Big\}d\gamma<\infty.$$

Making the transformation $u_j=\mu_j/\sqrt{(n_j-1)S_j^2+2b_j\gamma}$ gives

$$\int_{-\infty}^\infty\frac{\gamma^{a_j}\mu_j^2}{\psi_i(\mu_j,\gamma)^{n_j/2+a_j}}d\mu_j = \frac{\gamma^{a_j}}{\{(n_j-1)S_i^2+2b_j\gamma\}^{(n_j-3)/2+a_j}}\int_{-\infty}^\infty\frac{u_j^2}{(1+u_j^2)^{n_j/2+a_j}}du_j,$$

which is finite since $n_j>1$. Hence, (6.11) holds if

$$\int_0^\infty\left\{(n_*-1)S_*^2+2b_*\gamma\right\}^{-K/2}d\gamma<\infty,$$

where $K=n_j-3+\sum_{i\neq j}(n_i-1)=N-m-2$. This establishes that $E(\boldsymbol{\beta}\boldsymbol{\beta}^t|D)<\infty$ for $N>m+4$. Finally, for $E(\gamma^2|D)$, it follows that for $N>m+6$,

$$E(\gamma^2|D) < C\int_0^\infty\gamma^2\left\{\frac{1}{2}(n_*-1)S_*^2+b_*\gamma\right\}^{-(N-m)/2}d\gamma<\infty,$$

which completes the proof for (b).

6.5.2 Proof of Theorem 6.2.

We first prove part (a). From the proof of Theorem 6.1, it is sufficient to show that

$$\int_{\mathbb{R}_+\times\mathbb{R}^q}\prod_{i=1}^m\left(\gamma\exp(\boldsymbol{w}_i^t\boldsymbol{\eta})\right)^{a_i}\left\{\frac{1}{2}(n_i-1)S_i^2+b_i\gamma\exp(\boldsymbol{w}_i^t\boldsymbol{\eta})\right\}^{-(n_i/2+a_i)}d\gamma d\boldsymbol{\eta}<\infty, \tag{6.12}$$

under the condition that $t_k=1$ for $k=1,\ldots,q$. Since $(n_i-1)S_i^2$ and $\gamma\exp(\boldsymbol{w}_i^t\boldsymbol{\eta})$ are positive, the left side in (6.12) is evaluated from the upper by

$$\int_{\mathbb{R}_+\times\mathbb{R}^q}\gamma^A\prod_{k=1}^q\exp(\eta_k)^{B_{1k}}\Big\{C_*+b_*\gamma^{A+N/2+m}\prod_{k=1}^q\exp(\eta_k)^{B_{1k}+B_{2k}}\Big\}^{-1}d\gamma d\boldsymbol{\eta}, \tag{6.13}$$

where $A=\sum_{i=1}^m a_i$, $B_{1k}=\sum_{i=1}^m a_iw_{ik}$, $B_{2k}=2^{-1}\sum_{i=1}^m n_iw_{ik}$, $b_*=\prod_{i=1}^m b_i^{-(n_i/2+a_i)}$, and $C_*=2^{-(A+N/2+m)}\prod_{i=1}^m\{(n_i-1)S_i^2\}^{-(n_i/2+a_i)}$. Thus we need to show that (6.13) is finite. Without loss of generality, we consider the case of $B_{1k}>0$ and $B_{2k}>0$ for $k=1,\ldots,q$, since the case that $B_{1k}<0$ and $B_{2k}<0$ for some k reduces to $B_{1k}>0$ and $B_{2k}>0$ by changing the variable η_k as $-\eta_k$. From the positivity of B_{1k}'s, there exists $\lambda>0$ such that $B_{1k}>1/\lambda>0$ for $k=1,\ldots,q$, and we change the variables as $\phi_k=\exp(\eta_k/\lambda)$ in (6.13) to get $\int_{\mathbb{R}_+^{q+1}}f(\gamma,\boldsymbol{\phi})d\gamma d\boldsymbol{\phi}$, where $\boldsymbol{\phi}=(\phi_1,\ldots,\phi_q)$ and

$$f(\gamma,\boldsymbol{\phi})=\lambda^q\gamma^A\prod_{k=1}^q\phi_k^{\lambda B_{1k}-1}\Big(C_*+b_*\gamma^{A+N/2+m}\prod_{k=1}^q\phi_k^{\lambda B_{1k}+\lambda B_{2k}}\Big)^{-1}.$$

We decompose the integral $\int_{\mathbb{R}_+^{q+1}}f(\gamma,\boldsymbol{\phi})d\gamma d\boldsymbol{\phi}$ into the 2^{q+1} domains $\gamma\le 1$ or $\gamma\ge 1$, and $\phi_k\le 1$ or $\phi_k\ge 1$ for $k=1,\ldots,q$. Then it is sufficient to show that

$$\int_0^1\int_{(0,1]^r\times[1,\infty)^{q-r}}f(\gamma,\boldsymbol{\phi})d\boldsymbol{\phi}d\gamma<\infty,\qquad\int_1^\infty\int_{(0,1]^r\times[1,\infty)^{q-r}}f(\gamma,\boldsymbol{\phi})d\boldsymbol{\phi}d\gamma<\infty, \tag{6.14}$$

for fixed $r=0,\ldots,q$. For evaluating the former in (6.14), we define $g(\gamma,\phi_1,\ldots,\phi_r)=\int_{[1,\infty)^{q-r}}f(\gamma,\boldsymbol{\phi})d\boldsymbol{\phi}$. We note that $g(\gamma,\phi_1,\ldots,\phi_r)$ is 0 when at least one among $\gamma,\phi_1,\ldots,\phi_r$ is 0, and $g(\gamma,\phi_1,\ldots,\phi_r)<\infty$ for other values since

$$\begin{aligned}g(\gamma,\phi_1,\ldots,\phi_r)&=\lambda^q\gamma^A\prod_{k=1}^r\phi_k^{\lambda B_{1k}-1}\int_{[1,\infty)^{q-r}}\prod_{k=r+1}^q\phi_k^{\lambda B_{1k}-1}\Big(C_*+D_*\prod_{k=r+1}^q\phi_k^{\lambda B_{1k}+\lambda B_{2k}}\Big)^{-1}d\phi_{r+1}\ldots d\phi_q\\&\le\lambda^q\gamma^A\prod_{k=1}^r\phi_k^{\lambda B_{1k}-1}D_*^{-1}\prod_{k=r+1}^q\int_1^\infty\phi_k^{-\lambda B_{2k}-1}d\phi_k<\infty,\end{aligned}$$

for $0<\gamma,\phi_1,\ldots,\phi_r\le 1$, where $D_*=b_*\gamma^{A+N/2}\prod_{k=1}^r\phi_k^{\lambda B_{1k}+\lambda B_{2k}}$. Therefore, $g(\gamma,\phi_1,\ldots,\phi_r)$ is bounded over $[0,1]^r$, so that the former integral in (6.14) is finite. For the latter case of (6.14), we can similarly show that the integral is finite since $N/2>1$, which completes the proof for part (a).

For part (b), we first note that it can be proved of finiteness of the posterior variances of other parameters using the similar argument given in the proof of part (a) in Theorem 6.2. Hence, we show $E[\boldsymbol{\eta}_k^2|D], k = 1, \ldots, q$ are finite. To this end, it is sufficient to prove that

$$\int_{\mathbb{R}_+\times\mathbb{R}^q} \gamma^A \eta_k^2 \prod_{\ell=1}^q \exp(\eta_\ell)^{B_{1\ell}} \Big\{ C_* + b_* \gamma^{A+N/2} \prod_{k=1}^q \exp(\eta_k)^{B_{1\ell}+B_{2\ell}} \Big\}^{-1} d\gamma d\boldsymbol{\eta} < \infty,$$

for $k = 1, \ldots, q$. Under the condition that $B_{1k} > 0$ and $B_{2k} > 0$ for $k = 1, \ldots, q$, there exists $\lambda > 0$ such that $B_{1k} > 3/\lambda$ and $B_{2k} > 3/\lambda$, and we change the variables as $\phi_k = \exp(\eta_k/\lambda)$ in the left side to get $\int_{\mathbb{R}_+^{q+1}} f_k(\gamma, \boldsymbol{\phi}) d\gamma d\boldsymbol{\phi}$, where

$$f_k(\gamma, \boldsymbol{\phi}) = \lambda^3 \gamma^A \prod_{\ell=1}^q (\log \phi_k)^2 \phi_\ell^{\lambda B_{1\ell}-1} \Big\{ C_* + b_* \gamma^{A+N/2} \prod_{\ell=1}^q \phi_\ell^{\lambda B_{1\ell}+\lambda B_{2\ell}} \Big\}^{-1}.$$

We again decompose the 2^{q+1} domains $\gamma \le 1$ or $\gamma \ge 1$, and $\phi_k \le 1$ or $\phi_k \ge 1$ for $k = 1, \ldots, q$. Since $\lambda B_{1\ell} - 1 > 2$, $(\log \phi_k)^2 \phi_k^{\lambda B_{1k}-1}$ is bounded over $0 < \phi_k \le 1$. On the other hand, it is noted that $\int_1^\infty (\log \phi_k)^2 \phi_k^{\lambda B_{1k}-1}/(C + D\phi_k^{\lambda B_{1k}+\lambda B_{2k}}) d\phi_k = \int_0^\infty u^2 \exp(\lambda B_{1k} u)/(C + D\exp\{(\lambda B_{1k} + \lambda B_{2k})u\}) du < \infty$ under $B_{2k} > 0$. Therefore, similar evaluation shows that the integral $\int_{\mathbb{R}_+^{q+1}} f_k(\gamma, \boldsymbol{\phi}) d\gamma d\boldsymbol{\phi}$ is finite, whereby we complete the proof for part (b).

Chapter 7

Uncertain Random Effects

7.1 Introduction

Datta et al. (2011b) suggested inference by testing the presence of random effects in general mixed models. They pointed out that if the random effects can be dispensed with, the model parameters and the small area means may be estimated with substantially higher accuracy. Further, Datta and Mandal (2015) generalized the idea of preliminary testing to the uncertain random effects in the Fay-Herriot model, which assumes that, for all $i \in \{1, \ldots, m\}$,

$$y_i = \theta_i + \varepsilon_i, \quad \theta_i = \boldsymbol{x}_i^t \boldsymbol{\beta} + u_i v_i,$$

where $\varepsilon_i \sim N(0, D_i)$ for known D_i, $v_i \sim N(0, A)$ and $\Pr(u_i = 1) = p = 1 - \Pr(u_i = 0)$. In Datta and Mandal (2015), the term $u_i v_i$ is called the "uncertain random effect" since the density of $u_i v_i$ is expressed as a mixture of $N(0, A)$ and a point mass at 0. Because the distribution of the random effects is a mixture, the extent of these random effects can be controlled and flexible prediction can be achieved. Actually, the resulting estimator (predictor) of θ_i is expressed as the linear combination of the direct estimator y_i and the regression estimator $\boldsymbol{x}_i^t \widehat{\boldsymbol{\beta}}$. The weight depends on the squared residuals $(y_i - \boldsymbol{x}_i^t \widehat{\boldsymbol{\beta}})^2$ while the weight in the resulting estimator from the traditional Fay-Herriot model does not take the residuals into account.

In Datta and Mandal (2015), the Bayesian method was implemented for inferences of the small area parameters θ_i's as well as the model parameters by setting the proper prior distributions for p and A, namely $p \sim B(a_1, a_2)$ and $A \sim IG(a_3, a_4)$ for known (user specified) a_i, $i = 1, 2, 3, 4$, and the improper uniform prior for $\boldsymbol{\beta}$, where $B(a_1, a_2)$ and $IG(a_3, a_4)$ denote the beta and inverse gamma distribution, respectively. It was shown that the resulting posterior distributions of all the parameters are proper under some conditions. However, Datta and Mandal (2015) focused on the Fay–Herriot model, and their method could be restrictive in real applications. Moreover, they used a proper (informative) prior distribution for both p and A, and the result could be affected by the choice of hyperparameters.

In this chapter, we treat not only the uncertain random effects in more general small area models like the NER model, but also non-informative prior distributions for model parameters. The NER model has been used in various applications including small area estimation, biological experiments and econometric analysis. The NER model assumes that, for all

$i \in \{1, \ldots, m\}$ and $j \in \{1, \ldots, n_i\}$,

$$y_{ij} = \boldsymbol{x}_{ij}^t \boldsymbol{\beta} + v_i + \varepsilon_{ij},$$

where ε_{ij} is the sampling error associated with y_{ij} and v_i is a random effect in the ith area. It is usually assumed that ε_{ij} and v_i are mutually independent and distributed as $\varepsilon_{ij} \sim N(0, \sigma^2)$ and $v_i \sim N(0, \tau^2)$, respectively. The main purpose of the NER model is to predict (estimate) the quantity of linear combinations of $\boldsymbol{\beta}$ and v_i, namely $\mu_i = \boldsymbol{c}_i^t \boldsymbol{\beta} + v_i$ for some known vector $\boldsymbol{c}_i$.

In this chapter, we suggest the use of the uncertain random effect in the NER model and propose the uncertain nested error regression (UNER) model by adopting the structure

$$v_i | u_i \sim N(0, u_i \tau^2) \quad \text{with} \quad \Pr(u_i = 1) = p.$$

For the prior on τ^2, the variance of random effects, we use a distribution depending on the u_i's, which is defined as

$$\pi(\tau^2 | z > a) \propto \tau^{-1}, \quad \pi(\tau^2 | z \leq a) \propto \pi_*(\tau^2),$$

for some $a > 0$, where $z = u_1 + \cdots + u_m$ and $\pi_*(\tau^2)$ is some proper density, so that the prior distribution of τ^2 is more non-informative than the proper prior such as an inverse gamma distribution as used in Datta and Mandal (2015). For the other parameters $\boldsymbol{\beta}, \sigma^2$ and p, we also assign the non-informative prior $\pi(\boldsymbol{\beta}, \sigma^2, p) \propto p^{-1/2}(1-p)^{-1/2}\sigma^{-1}$. Hence, our Bayesian procedure is objective. We also apply the NER model in the framework of the finite population to predict the true finite population mean based on the partially observed data in each population.

This article is organized as follows. In Section 7.2, we describe the details of the UNER model and provide the full Bayesian method as well as the main theorem regarding the propriety of the posterior distribution and the finiteness of posterior variances. The prediction problem of finite population means using UNER is also discussed. In Section 7.3, we compare the UNER model with the NER model through simulation and empirical studies. The proof of the main result is given in Section 7.4.

7.2 Uncertain Nested Error Regression Models

7.2.1 Model settings and Bayes estimator

We consider the following uncertain nested error regression (UNER) model

$$\begin{aligned} &y_{ij} = \boldsymbol{x}_{ij}^t \boldsymbol{\beta} + v_i + \varepsilon_{ij}, \quad j = 1, \ldots, n_i, \\ &v_i | (u_i = 1) \sim N(0, \tau^2), \quad v_i | (u_i = 0) \sim \delta_0(v_i), \quad i = 1, \ldots, m, \end{aligned} \tag{7.1}$$

independently for i with $\Pr(u_i = 1) = 1 - \Pr(u_i = 0) = p$, where $\boldsymbol{x}_{ij}$ is a q-dimensional vector of covariates, $\boldsymbol{\beta}$ is a q-dimensional vector of regression coefficients, $\delta_0(\cdot)$ denotes the Dirac measure at 0, and the ε_{ij}'s are independently and identically distributed as $N(0, \sigma^2)$. The marginal density function of v_i is given by

$$f(v) = \frac{p}{\sqrt{2\pi}\tau} \exp\Big(-\frac{v^2}{2\tau^2} \Big) + (1-p) I(v = 0),$$

which is a mixture of the normal distribution $N(0, \tau^2)$ and the point mass at 0. Thus the model parameters are the regression coefficients $\boldsymbol{\beta}$, the variance components σ^2 and τ^2, and the mixture ratio p.

Let $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})^t$ be the observed vector in the ith area. Then the variance of $\boldsymbol{y}_i$ is $\mathrm{Var}(\boldsymbol{y}_i) = \sigma^2 \boldsymbol{I}_{n_i} + p\tau^2 \boldsymbol{J}_{n_i}$ for $\boldsymbol{J}_{n_i} = \mathbf{1}_{n_i}\mathbf{1}_{n_i}^t$. If the prior probability p of $u_i = 1$ is 0, it follows that $\mathrm{Var}(\boldsymbol{y}_i) = \sigma^2 \boldsymbol{I}_{n_i}$, and the observations in the ith area are mutually independent. The parameter that we want to estimate (predict) is $\mu_i = \boldsymbol{c}_i^t\boldsymbol{\beta} + v_i$ for a known vector $\boldsymbol{c}_i$. The typical choice of $\boldsymbol{c}_i$ is $\bar{\boldsymbol{x}}_i = (x_{i1} + \cdots + x_{in_i})/n_i$ in which μ_i corresponds to the mean of the ith area. The posterior distribution of μ_i given u_i and $\boldsymbol{y}_i$ is

$$\mu_i | u_i, \boldsymbol{y}_i \sim N\Big(\boldsymbol{c}_i^t\boldsymbol{\beta} + \frac{n_i\tau^2 I(u_i = 1)}{\sigma^2 + n_i\tau^2}(\bar{y}_i - \bar{\boldsymbol{x}}_i^t\boldsymbol{\beta}),\ \frac{I(u_i = 1)\sigma^2\tau^2}{\sigma^2 + n_i\tau^2}\Big),$$

where $\bar{y}_i = (y_{i1} + \cdots + y_{in_i})/n_i$, the sample mean of y_{ij} in the ith area. Thus the posterior distribution of μ_i given $\boldsymbol{y}_i$ is a mixture of the normal distribution and a point mass at $\boldsymbol{c}_i^t\boldsymbol{\beta}$. The resulting Bayes estimator $\widetilde{\mu}_i$ of μ_i is

$$\begin{aligned}\widetilde{\mu}_i = \mathrm{E}(\mu_i | \boldsymbol{y}_i) &= \widetilde{p}_i\Big\{\boldsymbol{c}_i^t\boldsymbol{\beta} + \frac{n_i\tau^2}{\sigma^2 + n_i\tau^2}(\bar{y}_i - \bar{\boldsymbol{x}}_i^t\boldsymbol{\beta})\Big\} + (1 - \widetilde{p}_i)\boldsymbol{c}_i^t\boldsymbol{\beta} \\ &= \boldsymbol{c}_i^t\boldsymbol{\beta} + \frac{n_i\tau^2\widetilde{p}_i}{\sigma^2 + n_i\tau^2}(\bar{y}_i - \bar{\boldsymbol{x}}_i^t\boldsymbol{\beta}),\end{aligned}$$

where $\widetilde{p}_i$ is the posterior probability of $u_i = 1$ given by

$$\begin{aligned}\widetilde{p}_i &= \Pr(u_i = 1 | \boldsymbol{y}_i) \\ &= p\Big[p + (1 - p)\sqrt{\frac{\sigma^2 + n_i\tau^2}{\sigma^2}}\exp\Big\{-\frac{n_i^2\tau^2}{2\sigma^2(\sigma^2 + n_i\tau^2)}(\bar{y}_i - \bar{\boldsymbol{x}}_i^t\boldsymbol{\beta})^2\Big\}\Big]^{-1}.\end{aligned} \tag{7.2}$$

We note that $\widetilde{p}_i$ increases in p and $(\bar{y}_i - \bar{\boldsymbol{x}}_i^t\boldsymbol{\beta})^2$. Thus, if $\boldsymbol{x}_{ij}$ is a good covariate to explain y_{ij} in the ith area, the squared residual $(\bar{y}_i - \bar{\boldsymbol{x}}_i^t\boldsymbol{\beta})^2$ is expected to be small, and the posterior probability $\widetilde{p}_i$ is small as well. The posterior probability $\widetilde{p}_i$ is 1 when $p = 1$ and $\widetilde{p}_i$ converges to 1 as $(\bar{y}_i - \bar{\boldsymbol{x}}_i^t\boldsymbol{\beta})^2$ goes to infinity.

Moreover, the posterior variance of μ_i is expressed as

$$\begin{aligned}V_i(\boldsymbol{y}_i) \equiv \mathrm{Var}(\mu_i | \boldsymbol{y}_i) &= \mathrm{Var}(v_i | \boldsymbol{y}_i) \\ &= \frac{n_i^2\tau^4}{(\sigma^2 + n_i\tau^2)^2}(\bar{y}_i - \bar{\boldsymbol{x}}_i^t\boldsymbol{\beta})^2\widetilde{p}_i(1 - \widetilde{p}_i) + \frac{\sigma^2\tau^2\widetilde{p}_i}{\sigma^2 + n_i\tau^2}.\end{aligned} \tag{7.3}$$

It is worth pointing out that in this case, the posterior variance of μ_i depends on the observation $\boldsymbol{y}_i$ through the squared residual $(\bar{y}_i - \bar{\boldsymbol{x}}_i^t\boldsymbol{\beta})^2$ and the posterior probability $\widetilde{p}_i$, while the posterior variance of the random effect in the usual nested error regression model is given by $\sigma^2\tau^2(\sigma^2 + n_i\tau^2)^{-1}$, which does not depend on $\boldsymbol{y}_i$. This means that the uncertain random effect enables us to take the distance between the sample mean $\bar{y}_i$ and the synthetic estimator $\bar{\boldsymbol{x}}_i^t\boldsymbol{\beta}$ into the posterior variability of the parameter of interest, μ_i.

7.2.2 Bayesian implementation and posterior distribution

Since the marginal likelihood function of the model parameters $\boldsymbol{\beta}, \sigma^2, \tau^2$ and p is rather complex, we consider objective Bayesian inference for the model parameters as well as the

random effect v_i. To this end, we rewrite the model (7.1) as

$$\begin{aligned} &y_{ij}|v_i, \boldsymbol{\beta}, \sigma^2 \sim N(\boldsymbol{x}_{ij}^t\boldsymbol{\beta} + v_i, \sigma^2), \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, m \\ &v_i|u_i, \tau^2 \sim N(0, u_i\tau^2), \quad u_i|p \sim \text{Bin}(1, p), \quad i = 1, \ldots, m \end{aligned} \tag{7.4}$$

independently for i, where $\text{Bin}(1, p)$ denotes the Bernoulli distribution. To implement a full Bayesian inference, we need to set prior distributions on the model parameters. To keep the inference objective, we use the uniform prior distribution on $\boldsymbol{\beta}$ and the Jeffreys prior distributions on σ^2 and p. On the other hand, the prior distribution of τ^2 should depend on $z = u_1 + \cdots + u_m$, since τ^2 cannot be identified for a small value of z. Thus, for the model parameters, we use the prior distributions

$$\pi(\boldsymbol{\beta}, \sigma^2, p) = p^{-1/2}(1-p)^{-1/2}\sigma^{-1}, \quad \pi(\tau^2|z) \propto \begin{cases} \tau^{-1} & \text{if } z > a \\ \pi_*(\tau^2) & \text{if } z \le a \end{cases} \tag{7.5}$$

where $\pi_*(\tau^2) = (\tau^2)^{-b_1-1}\exp(-b_2/\tau^2)$ for known constants $b_1 > 3$ and $b_2 > 0$. The value of a is chosen by the user, and this point will be discussed later. It is noted that the prior distribution on p is proper, but the priors on $\boldsymbol{\beta}, \sigma^2$ and τ^2 are improper, so that the posterior propriety is not always guaranteed. In Theorem 7.1, we show that the posterior distribution for the model parameters is proper under mild conditions.

We now describe the posterior distribution and investigate its properties. The set of all observed data is denoted by $D = \{\boldsymbol{y}_i, \boldsymbol{X}_i\}_{i=1,\ldots,m}$ for $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in_i})$. From the model (7.4) with prior setup (7.5), the posterior density of the parameters $(\boldsymbol{v}, \boldsymbol{u}, \boldsymbol{\beta}, \sigma^2, \tau^2, p)$ for $\boldsymbol{v} = (v_1, \ldots, v_m)^t$ and $\boldsymbol{u} = (u_1, \ldots, u_m)^t$ is given by

$$\begin{aligned} &\pi(\boldsymbol{v}, \boldsymbol{u}, \boldsymbol{\beta}, \sigma^2, \tau^2, p|D) \\ &\propto (\sigma^2)^{-(N+1)/2}(\tau^2)^{-\{z+I(z>a)\}/2-(b_1+1)I(z\le a)}p^{z-1/2}(1-p)^{m-z-1/2} \\ &\quad \times \prod_{i=1}^{m}\Big[\exp\Big\{-\frac{\sum_{j=1}^{n_i}(y_{ij} - \boldsymbol{x}_{ij}^t\boldsymbol{\beta} - v_i)^2}{2\sigma^2} - \frac{u_i v_i^2}{2\tau^2}\Big\}\delta_0(v_i)^{1-u_i}\Big] \\ &\quad \times \exp\Big\{-\frac{b_2}{\tau^2}I(z \le a)\Big\}. \end{aligned} \tag{7.6}$$

We can now state our main result about the posterior propriety and the existence of posterior variances.

Theorem 7.1. *The following statements hold true.*

(a) *The marginal posterior density $\pi(\boldsymbol{\beta}, \sigma^2, \tau^2, p|D)$ is proper if $N > q + 2$ and $m > a \ge 1$.*

(b) *The model parameters $\boldsymbol{\beta}, \sigma^2, \tau^2$ and p have finite posterior variances if $N > q + 6$ and $m > a \ge 5$.*

Remember that q is the dimension of the vector of regression coefficients $\boldsymbol{\beta}$, and a is the tuning parameter of the prior for τ^2. Part (a) in Theorem 7.1 says that the marginal posterior densities of the small area means are proper and part (b) provides a sufficient condition for obtaining finite measures of uncertainty for the model parameters. We note that the conditions in Theorem 7.1 are similar to the conditions given in Arima et al. (2015) and Datta and Mandal (2015). The proof of Theorem 7.1 is presented in Section 7.4.

Since the posterior distribution in (7.6) cannot be obtained in closed form, we rely on the Markov chain Monte Carlo technique, in particular the Gibbs sampler, in order to draw samples from the posterior distribution. This requires generating samples from the full conditional distributions for each of $(\boldsymbol{v}, \boldsymbol{u}, \boldsymbol{\beta}, \sigma^2, \tau^2, p)$ given the remaining parameters and the data D. Fortunately, the full conditional distributions can be described using familiar distributions allowing us to easily implement the Gibbs sampling. The full conditional distributions are given, for all $i \in \{1, \ldots, m\}$, by

$$\begin{aligned}
&v_i|u_i, \boldsymbol{\beta}, \sigma^2, \tau^2, D \sim N\Big[\frac{n_i\tau^2 I(u_i=1)}{\sigma^2 + n_i\tau^2}(\bar{y}_i - \bar{\boldsymbol{x}}_i^t\boldsymbol{\beta}), \frac{\sigma^2\tau^2 I(u_i=1)}{\sigma^2 + n_i\tau^2}\Big],\\
&u_i|\boldsymbol{\beta}, \sigma^2, \tau^2, p, D \sim \text{Bin}(1, \widetilde{p}_i), \quad p|\boldsymbol{u}, D \sim B\Big(z + \frac{1}{2}, m - z + \frac{1}{2}\Big),\\
&\boldsymbol{\beta}|\boldsymbol{u}, \sigma^2, \tau^2, D \sim N_p[(\boldsymbol{X}^t\boldsymbol{\Sigma}_u^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{\Sigma}_u^{-1}\boldsymbol{y}, (\boldsymbol{X}^t\boldsymbol{\Sigma}_u^{-1}\boldsymbol{X})^{-1}],\\
&\tau^2|\boldsymbol{u}, \boldsymbol{v}, D \sim IG\Big[\frac{1}{2}\{z - I(z > a)\} + b_1 I(z \le a), \frac{1}{2}\sum_{i=1}^m u_i v_i^2 + b_2 I(z \le a)\Big],\\
&\sigma^2|\boldsymbol{v}, \boldsymbol{\beta}, D \sim IG\Big[\frac{1}{2}(N-1), \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{v})^t(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{v})\Big],
\end{aligned} \tag{7.7}$$

where

$$z = \sum_{i=1}^m u_i, \quad \boldsymbol{\Sigma}_u = \text{diag}(\boldsymbol{\Sigma}_{1u}, \ldots, \boldsymbol{\Sigma}_{mu})$$

with

$$\boldsymbol{\Sigma}_{iu} = \sigma^2\boldsymbol{I}_{n_i} + u_i\tau^2\mathbf{1}_{n_i}\mathbf{1}_{n_i}^t,$$

$\boldsymbol{y} = (\boldsymbol{y}_1^t, \ldots, \boldsymbol{y}_m^t)^t$, $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m)$, and $\widetilde{p}_i$ is given in (7.2). Using these expressions of full conditional distributions, we can easily draw posterior samples of all the variances and parameters to make inferences, such as point estimation, prediction intervals and standard errors, for $\mu_i = \boldsymbol{c}_i^t\boldsymbol{\beta} + v_i$.

In closing of this section, we discuss the choices of a, b_1 and b_2 in the posterior distribution of τ^2. Remember that the prior distribution of τ^2 is non-informative and improper when $z > a$ and informative and proper when $z \le a$. Taking this into account, we should select a value of a as small as possible. Hence, it follows from Theorem 7.1 that $a = 5$ is the most reasonable choice. On the other hand, as discussed in Datta and Mandal (2015), a reasonable choice is $b_1 = V + 2$ and $b_2 = V(V+1)$ such that $\text{E}(\tau^2|z \le a) = V$ and $\text{Var}(\tau^2|z \le a) = V^2$, where V is the estimated sampling variance given by

$$V = \frac{1}{N - m - q}\sum_{i=1}^m\sum_{j=1}^{n_i}\{y_{ij} - \bar{y}_i - (\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}})^t\widehat{\boldsymbol{\beta}}_{\text{OLS}}\}^2.$$

Here, $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$ is the ordinary least squared estimator of $\boldsymbol{\beta}$. It should be noted that V satisfies $\text{E}(V) = \sigma^2$.

7.2.3 Prediction in finite populations

Here, we consider the problem of predicting the means in finite populations. Assume that there exist m finite populations and the ith population consists of N_i pairs of data $(Y_{i1}, \boldsymbol{x}_{i1}), \ldots, (Y_{iN_i}, \boldsymbol{x}_{iN_i})$.

It is supposed that n_i ($< N_i$) observations are sampled from the ith population. What we want to predict is the mean of the ith finite population $\bar{Y}_i = (Y_{i1} + \cdots + Y_{iN_i})/N_i$. Assume also that the mean vector of covariates $\bar{\boldsymbol{X}}_i = (\boldsymbol{x}_{i1} + \cdots + \boldsymbol{x}_{iN_i})/N_i$ is available, which is often encountered in real application, see Battese et al. (1988).

Let s_i and r_i be collections of indices of sampled and non-sampled observations in the ith area, respectively, so that s_i and r_i satisfy $s_i \cap r_i = \phi$ and $s_i \cup r_i = \{1, \ldots N_i\}$. Without loss of generality, we assume that $s_i = \{1, \ldots, n_i\}$ and $r_i = \{n_i + 1, \ldots, N_i\}$. The Bayes estimator of $\bar{Y}_i$ under quadratic loss is given by

$$\mathrm{E}(\bar{Y}_i|\boldsymbol{y}_i) = \frac{1}{N_i}\Big\{ n_i \bar{y}_{i(s)} + (N_i - n_i)\mathrm{E}(\bar{Y}_{i(r)}|\boldsymbol{y}_i)\Big\},$$

where

$$\bar{y}_{i(s)} = \frac{1}{n_i}\sum_{j \in s_i} y_{ij}, \quad \bar{Y}_{i(r)} = \frac{1}{N_i - n_i}\sum_{j \in r_i} Y_{ij}.$$

For evaluating the conditional expectation $\mathrm{E}(\bar{Y}_{i(r)}|\boldsymbol{y}_i)$, we assume that Y_{ij} is expressed, for each $j \in r_i$, by

$$Y_{ij} = \boldsymbol{x}_{ij}^t \boldsymbol{\beta} + v_i + \varepsilon_{ij},$$

that is, the non-sampled observations have the same data generating structure as the sampled ones. Then the unobserved mean $\bar{Y}_{i(r)}$ is expressed as

$$\bar{Y}_{i(r)} = \bar{\boldsymbol{x}}_{i(r)}^t \boldsymbol{\beta} + v_i + \bar{\varepsilon}_{i(r)},$$

where

$$\bar{\varepsilon}_{i(r)} = \frac{1}{N_i - n_i}\sum_{j \in r_i} \varepsilon_{ij}.$$

Thus the conditional distribution of $\bar{Y}_{i(r)}$ given $\boldsymbol{y}_i$ and u_i is

$$\bar{Y}_{i(r)}|\boldsymbol{y}_i, u_i \sim N\Big[\bar{\boldsymbol{x}}_{i(r)}^t \boldsymbol{\beta} + \frac{I(u_i = 1) n_i \tau^2}{\sigma^2 + n_i \tau^2}(\bar{y}_i - \bar{\boldsymbol{x}}_i^t \boldsymbol{\beta}),\ \frac{I(u_i = 1)\sigma^2 \tau^2}{\sigma^2 + n_i \tau^2} + \frac{\sigma^2}{N_i - n_i}\Big], \tag{7.8}$$

which yields the predictive density of $\bar{Y}_{i(r)}$ given by

$$\begin{aligned}\bar{Y}_{i(r)}|\boldsymbol{y}_i \sim\ & \widetilde{p}_i\, N\Big[\bar{\boldsymbol{x}}_{i(r)}^t \boldsymbol{\beta} + \frac{n_i \tau^2}{\sigma^2 + n_i \tau^2}(\bar{y}_i - \bar{\boldsymbol{x}}_i^t \boldsymbol{\beta}),\ \frac{\sigma^2 \tau^2}{\sigma^2 + n_i \tau^2} + \frac{\sigma^2}{N_i - n_i}\Big] \\ & + (1 - \widetilde{p}_i)\, N\Big[\bar{\boldsymbol{x}}_{i(r)}^t \boldsymbol{\beta}, \frac{\sigma^2}{N_i - n_i}\Big],\end{aligned}$$

where $\widetilde{p}_i$ is the posterior probability of $u_i = 1$ given in (7.2). Thus the conditional distribution of the non-sampled data is a mixture of the two normal distributions of the predictive density, with and without random effect. Moreover, the conditional variance $\bar{Y}_{i(r)}$ given $\boldsymbol{y}_i$ is calculated as $V_i(\boldsymbol{y}_i) + \sigma^2/(N_i - n_i)$, where $V_i(\boldsymbol{y}_i)$ is the posterior variance of v_i given in (7.3). It is noted that, when the true mean vector of the explanatory variables $\bar{\boldsymbol{X}}_i$ is available in each area, the value of $\bar{\boldsymbol{x}}_{i(r)}$ is easily obtained by

$$\bar{\boldsymbol{x}}_{i(r)} = \frac{1}{N_i - n_i}(N_i \bar{\boldsymbol{X}}_i - n_i \bar{\boldsymbol{x}}_i).$$

To implement the prediction in the finite population model, we regard the $\bar{Y}_{i(r)}$'s as latent variables and add the sampling step from (7.8) to the Gibbs sampling given in (7.7).

7.3 Numerical studies

7.3.1 Model-based simulations

In this simulation study, we compared the UNER model with the conventional NER model in terms of the quality of the estimates. In applying the NER model, we used the Jeffreys prior on $(\boldsymbol{\beta}, \tau^2, \sigma^2)$, namely $\pi(\boldsymbol{\beta}, \tau^2, \sigma^2) = \tau^{-1}\sigma^{-1}$, where it is well-known that the resulting posterior distribution is proper; see Berger (1985). The full conditional posterior distributions are given by

$$\begin{aligned}
&v_i|\boldsymbol{\beta}, \sigma^2, \tau^2, D \sim N\Big[\frac{n_i\tau^2}{\sigma^2 + n_i\tau^2}(\bar{y}_i - \bar{\boldsymbol{x}}_i^t\boldsymbol{\beta}), \frac{\sigma^2\tau^2}{\sigma^2 + n_i\tau^2}\Big], \quad i = 1, \ldots, m \\
&\boldsymbol{\beta}|\tau^2\sigma^2, D \sim N_p[(\boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{y}, (\boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}], \\
&\tau^2|\boldsymbol{v}, D \sim IG\Big(\frac{1}{2}(m-1), \frac{1}{2}\sum_{i=1}^{m} v_i^2\Big), \\
&\sigma^2|\boldsymbol{v}, \boldsymbol{\beta}, D \sim IG\Big[\frac{1}{2}(N-1), \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{v})^t(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{v})\Big],
\end{aligned} \tag{7.9}$$

where $\boldsymbol{\Sigma} = \text{diag}(\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_m)$ with $\boldsymbol{\Sigma}_i = \sigma^2\boldsymbol{I}_{n_i} + \tau^2\mathbf{1}_{n_i}\mathbf{1}_{n_i}^t$. We considered the following data generating process: for all $j \in \{1, \ldots, n\}$ and $i \in \{1, \ldots, m\}$,

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + v_i + \varepsilon_{ij},$$

where $\varepsilon_{ij} \sim N(0, 1)$, $\beta_0 = 1$, $\beta_1 = 0.5$, and the x_{ij}'s are $\mathcal{U}(1, 2)$ and fixed through simulation runs. Four combinations of (n, m) were considered, namely $(n, m) = (5, 20)$, $(5, 40)$, $(10, 20)$, $(10, 40)$. For the true distributions of v_i, we considered the following four scenarios for each choice of (n, m), viz.

$$\begin{aligned}
&\text{S1: } v_i \sim N(0, (0.7)^2), \quad \text{S2: } v_i \sim 0.3\delta_0(v_i) + 0.7N[0, (0.7)^2], \\
&\text{S3: } v_i \sim 0.3\delta_0(v_i) + 0.7\mathcal{L}[0, (0.7)^2], \quad \text{S4: } v_i \sim 0.3\delta_0(v_i) + 0.7t_6[0, (0.7)^2],
\end{aligned}$$

where $t_6(a, b)$ and $\mathcal{L}(a, b)$ denote the scaled t-distribution with 6 degrees of freedom with mean a and variance b and the scaled Laplace distribution with mean a and variance b, respectively. Hence, UNER is misspecified in scenarios S3 and S4, and over-specified in scenario S1.

Based on $R = 1000$ simulation runs, we computed the mean squared errors (MSE), absolute bias of $\widehat{\mu}_i$, and empirical coverage probability of the 95% credible interval of μ_i, which are respectively defined as

$$\begin{aligned}
&\text{MSE} = \frac{1}{mR}\sum_{r=1}^{R}\sum_{i=1}^{m}(\widehat{\mu}_i^{(r)} - \mu_i^{(r)})^2, \quad \text{Bias} = \frac{1}{mR}\sum_{r=1}^{R}\sum_{i=1}^{m}|\widehat{\mu}_i^{(r)} - \mu_i^{(r)}| \\
&\text{CP} = \frac{1}{mR}\sum_{r=1}^{R}\sum_{i=1}^{m} I(\mu_i^{(r)} \in \text{CI}_i^{(r)}) \times 100,
\end{aligned}$$

where $\widehat{\mu}_i^{(r)}$, $\mu_i^{(r)}$ and $\text{CI}_i^{(r)}$ are the posterior mean, the true value, and the 95% credible intervals, respectively, of μ_i in the rth simulation run. In each iteration of the simulation run, we used 5000 posterior samples after 1000 initial iterations for both UNER and NER.

Table 7.1: Simulated MSE, Bias and Coverage Probabilities (CP) of UNER and NER in Different Scenarios.

		UNER			NER		
(n, m)	Scenario	MSE	Bias	CP	MSE	Bias	CP
$(3, 25)$	S1	0.278	0.419	92.3	0.265	0.408	92.3
	S2	0.165	0.308	93.6	0.176	0.320	93.4
	S3	0.156	0.293	93.3	0.166	0.309	93.2
	S4	0.163	0.301	93.9	0.172	0.313	93.8
$(3, 50)$	S1	0.248	0.396	93.2	0.242	0.388	93.2
	S2	0.126	0.252	94.3	0.136	0.267	94.3
	S3	0.128	0.245	93.6	0.140	0.261	93.6
	S4	0.130	0.258	94.6	0.140	0.272	94.3
$(6, 25)$	S1	0.160	0.319	93.7	0.154	0.313	93.7
	S2	0.088	0.215	94.1	0.098	0.235	94.1
	S3	0.088	0.217	93.7	0.103	0.239	93.7
	S4	0.094	0.221	93.8	0.104	0.240	93.8
$(6, 50)$	S1	0.144	0.302	94.3	0.141	0.299	94.3
	S2	0.076	0.206	94.5	0.095	0.229	94.5
	S3	0.071	0.180	94.3	0.091	0.216	94.3
	S4	0.077	0.191	95.1	0.088	0.216	95.1

The results are given in Table 7.1. In scenario S1, both the MSE and absolute bias of UNER are larger than those of NER since UNER is over-specified. However, as the number of n and m get large, the difference of these values gets small. For the other scenarios, we can observe that UNER clearly performs better than NER in terms of MSE and absolute bias, and the differences get larger as n and m get larger. Finally, it is observed that the coverage probability of credible intervals are similar in UNER and NER. Hence, we can conclude that UNER is expected to be a useful tool when m and n are moderate or large.

7.3.2 Application to PLP data in Japan

This example, primarily for illustration, we used the UNER model (7.1) and data from the posted land price data along the Keikyu train line in 2001, which were treated in Section 5.4.3. For all $j \in \{1, \ldots, n_i\}$, let y_{ij} denote the log-transformed value of the posted land price (Yen) per for square meter of the jth spot, T_i is the time it takes from the nearby station i to Tokyo station around 8:30 in the morning, D_{ij} is the geographical distance from spot j to station i and FAR_{ij} denotes the floor-area ratio, or ratio of the building volume to the area at spot j. These values of T_i, D_{ij} and FAR_{ij} are also transformed by the logarithmic function. We applied the following UNER model:

$$\begin{aligned} &y_{ij} = \beta_0 + FAR_{ij}\beta_1 + T_i\beta_2 + D_{ij}\beta_3 + v_i + \varepsilon_{ij}, \\ &v_i|(u_i = 1) \sim N(0, \tau^2), \quad v_i|(u_i = 0) \sim \delta_0(v_i), \end{aligned} \tag{7.10}$$

where the ε_{ij}'s are independent and identically distributed as $N(0, \sigma^2)$. For comparison, we also applied the conventional NER model to this data set.

In applying the UNER model, we used the prior distribution with $a = 5$ and $b_1 = V + 2, b_2 = V(V+1)$ for $V = 0.031$ as discussed in the end of Section 7.2.2. In both models, we generated 100000 posterior samples after 10000 iterations of Gibbs sampling given in (7.7) and (7.9), respectively, and obtained the posterior means as well as the 95% credible intervals of the model parameters, which are given in Table 7.2. Moreover, based on the posterior samples, we computed the Deviance Information Criterion (DIC) suggested by Spiegelhalter et al. (2002), which is defined as

$$\mathrm{DIC} = 2\overline{D(\boldsymbol{\phi})} - D(\overline{\boldsymbol{\phi}}),$$

where $\boldsymbol{\phi}$ is a vector of the unknown model parameters, $D(\boldsymbol{\phi})$ is (-2) times the log-marginal likelihood function, and $\overline{D(\boldsymbol{\phi})}$ and $\overline{\boldsymbol{\phi}}$ denote the posterior means of $D(\boldsymbol{\phi})$ and $\boldsymbol{\phi}$, respectively. Note that $\boldsymbol{\phi} = \{\boldsymbol{\beta}, \tau^2, \sigma^2, p\}$ in the UNER model, and $\boldsymbol{\phi} = \{\boldsymbol{\beta}, \tau^2, \sigma^2\}$ in the NER model, which are given in Table 7.2 as well.

Table 7.2 shows that the posterior estimates and credible intervals of the regression coefficients $\beta_1, \ldots, \beta_4$ are similar between UNER and NER, and in both models, all the credible intervals of the regression coefficients are bounded away from 0. On the other hand, the results for the variance components τ^2 and σ^2 are different because of the effect of the parameter p. In terms of DIC values, the UNER model seems preferable to the conventional NER model.

To see the effects of u_i, we computed the posterior probabilities $\widetilde{p}_i$'s which are illustrated in the left panel in Figure 7.1. It is apparent that the $\widetilde{p}_i$'s change dramatically from area to area, and the $\widetilde{p}_i$'s in most areas are around 0.5, which comes from the posterior mean of $p = 0.54$ as shown in Table 7.2.

We next considered estimating the land price of a spot with a floor-area ratio of 100% and a distance of 1000m from the station i, namely

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

for $FAR_0 = \log(100)$ and $D_0 = \log(1000)$ in base 10. Based on the posterior samples, we calculated the point estimates $\widehat{\mu}_i$ and the posterior standard errors. The results are given in the right panel of Figure 7.1. Note that the mean of the posterior standard errors for all areas in UNER and NER are 6.5×10^{-2} and 6.8×10^{-2}, respectively.

We also computed the length of the prediction intervals of μ_i, and found that the results are similar to standard errors. It is clear from Figure 7.1 that UNER provides better estimates than NER in terms of posterior standard errors in most areas. In some areas, the posterior standard errors of UNER are larger than those of NER when correspondingly the posterior probability $\widetilde{p}_i$ is larger than 0.7 as shown in the left panel of Figure 7.1. Thus the uncertain random effects may increase the variability of predictors compared to the conventional random effects in the areas where the existence of random effect is strongly supported. This phenomenon was pointed out by Datta and Mandal (2015) for the Fay–Herriot model. However, taking the DIC values into account as well, the UNER model works well in this application.

7.3.3 Design-based simulation

We next investigated the numerical performance of the small area prediction problem in the framework of a finite population. We again used the PLP data in the Kanto region in 2001,

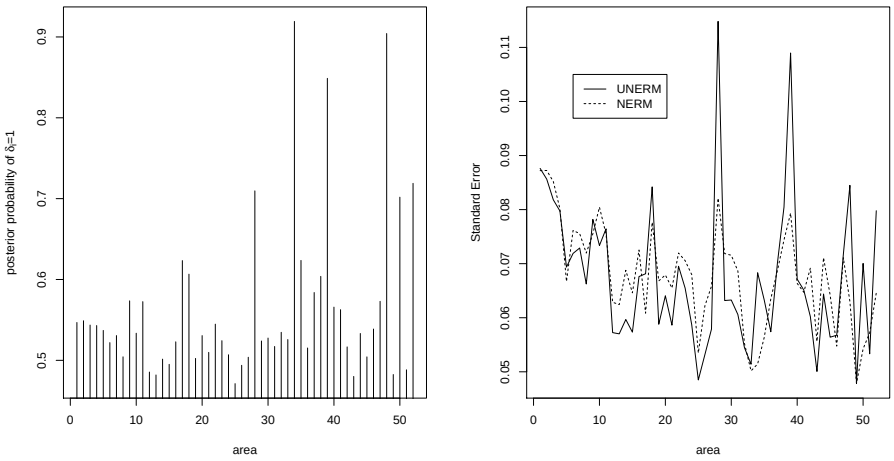

Figure 7.1: Posterior Probability of $u_i = 1$ (Left) and Standard Errors of μ_i (Right) in Each Area.

Table 7.2: Posterior Means and Credible Intervals of the Model Parameters, and DIC.

		β_0	β_1	β_2	β_3	σ^2	τ^2	p	DIC
	95%CI (upper)	15.16	0.24	-0.53	-0.051	0.041	0.071	0.99	
UNER	mean	14.55	0.17	-0.61	-0.091	0.033	0.017	0.54	512.6
	95%CI (lower)	13.88	0.11	-0.69	-0.131	0.026	0.002	0.05	
	95%CI (upper)	15.17	0.24	-0.53	-0.050	0.20	0.117	$-$	
NER	mean	14.52	0.17	-0.61	-0.089	0.18	0.075	$-$	703.1
	95%CI (lower)	13.88	0.10	-0.69	-0.132	0.16	0.031	$-$	

which includes the prefectures of Tokyo, Kanagawa, Chiba and Saitama. Thus the data set includes the PLP data along the Keikyu line used in the previous subsection. The full data set we used is the land price data with covariates (T_i, D_{ij} and FAR_{ij} as used in the previous study) and each data point has its unique nearest railroad station, which we regard as a small area.

For the ith small area ($i = 1, \ldots, m$), there are N_i land spots. To consider all the observed land price data in each small area in the framework of a finite population, we analyzed only the data which belong to the small areas that have a moderately large number of data points, namely we chose the areas i with $N_i \geq 20$. Then the resulting number of finite populations is $m = 30$, and the population sizes N_i's range from 20 to 45, but most N_i's vary around 25.

We artificially made the sampled data set and predict each finite population mean of the land price by applying UNER. The sampling scheme is simple random sampling without replacement in each finite population and n_i data are sampled in the ith finite population. The sample sizes n_i's are decided by some ratio $0 < \pi < 1$ and 100π percent of the data in each population are sampled, i.e., n_i is the nearest integer to $N_i \times \pi$. We considered four choices for π, namely $\pi = 0.3, 0.5, 0.7, 0.9$. In each case, we computed the squared root mean

squared errors for estimators of finite population means as

$$\text{SMSE}_i = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\widehat{\mu}_i^{(r)} - \mu_i)^2},$$

where $\widehat{\mu}_i^{(r)}$ is the estimator of the finite population using UNER or NER, and $R = 1000$ in this study. For both UNER and NER, we calculated $\widehat{\mu}_i^{(r)}$ by 5000 posterior samples after 1000 iterations using the method discussed in Section 7.2.3. In the UNER estimation, the same form of the prior distribution as in the previous section was used, namely $a = 5, b_1 = V + 2$ and $b_2 = V(V + 1)$ for estimated sampling error V.

To compare values of the SMSE for the two models, we then computed the ratio of SMSE given by $\text{SMSE}_i^{\text{UNER}}/\text{SMSE}_i^{\text{NER}}$, and provide their values in Figure 7.2. It is observed from Figure 7.2 that UNER provides better estimates than NER in some areas, but worse estimates than in several areas for the four cases of π. Moreover, it is also revealed that an improvement of UNER over NER becomes greater as the sampling rate π gets larger.

7.4 Proof of Theorem 7.1.

Let π^* be the right side of (7.6). For part (a), we shall show that

$$\sum_{\boldsymbol{u}\in\{0,1\}^m}\int \pi^*(\boldsymbol{v},\boldsymbol{u},\boldsymbol{\beta},\sigma^2,\tau^2,p|D)d\boldsymbol{v}d\boldsymbol{\beta}d\sigma^2d\tau^2dp < \infty,$$

namely the integral for each $\boldsymbol{u}$ is finite. We first prove for the case $\boldsymbol{u} = (0,\ldots,0)^t$. In this case, the integral reduces to

$$\int (\sigma^2)^{-(N+1)/2}(1-p)^{m-1/2}\exp\Big\{-\frac{1}{2\sigma^2}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij}-\boldsymbol{x}_{ij}^t\boldsymbol{\beta})^2\Big\}d\boldsymbol{\beta}d\sigma^2ddp.$$

Note that

$$\int p^{-1/2}(1-p)^{m-1/2}dp = B(1/2, m+1/2),$$

where $B(a,b)$ is a beta function. Then the integral is finite since the posterior distribution of the usual linear regression for the Jeffreys prior is proper if the conditions given in Theorem 7.1 are satisfied.

For the integral in the case $z \geq 1$, using $p^{z-1/2}(1-p)^{m-z-1/2} \leq 1$, it is sufficient to show that

$$\int \pi_u(\boldsymbol{v},\sigma^2,\tau^2,\boldsymbol{\beta})d\boldsymbol{v}d\boldsymbol{\beta}d\sigma^2d\tau^2 < \infty,$$

for

$$\pi_u(\boldsymbol{v},\sigma^2,\tau^2,\boldsymbol{\beta}) = \begin{cases}\pi_{u1}(\boldsymbol{v},\sigma^2,\tau^2,\boldsymbol{\beta}) & \text{if } z > a\\ \pi_{u2}(\boldsymbol{v},\sigma^2,\tau^2,\boldsymbol{\beta}) & \text{if } 0 < z \leq a\end{cases} \tag{7.11}$$

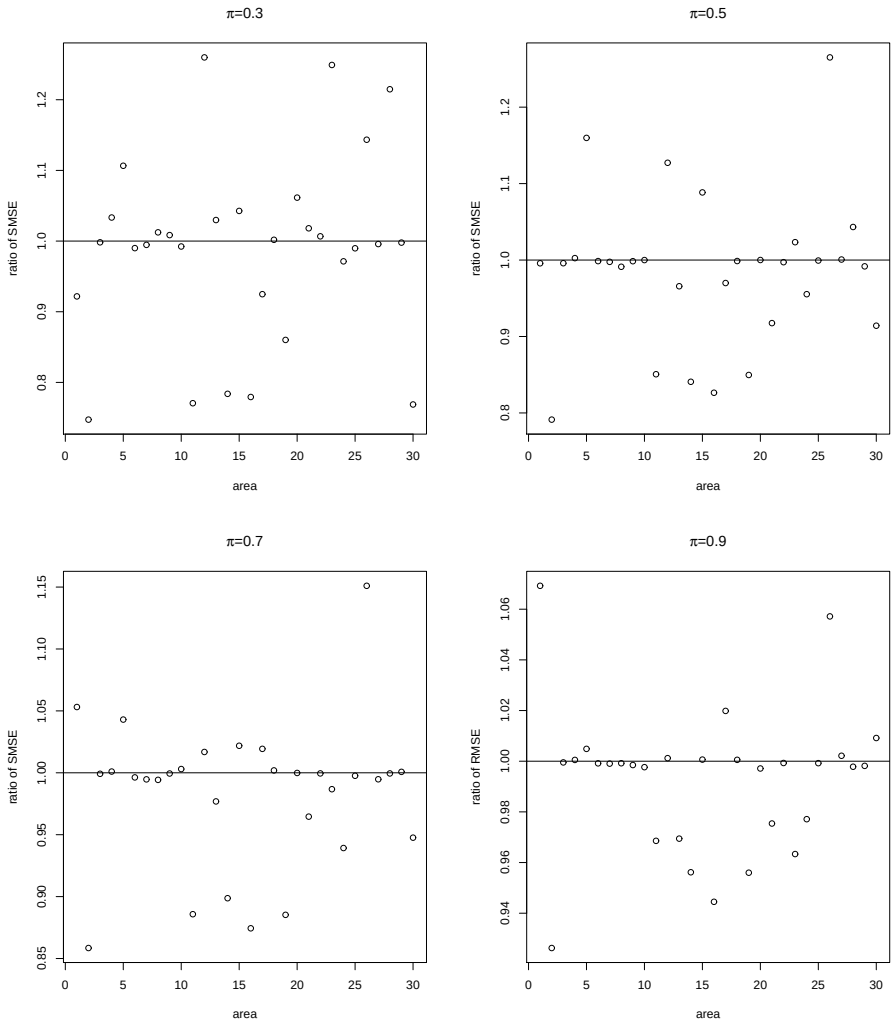

Figure 7.2: Squared Root Mean Squared Errors of Estimation of Finite Population Mean.

where

$$\pi_{u1}(\boldsymbol{v},\sigma^2,\tau^2,\boldsymbol{\beta}) = (\sigma^2)^{-(N+1)/2}(\tau^2)^{-(z+1)/2}\prod_{i=1}^{m}\delta_0(v_i)^{1-u_i} \times\prod_{i=1}^{m}\Big[\exp\Big\{-\frac{1}{2\sigma^2}\sum_{j=1}^{n_i}(y_{ij}-\boldsymbol{x}_{ij}^t\boldsymbol{\beta}-v_i)^2-\frac{u_i v_i^2}{2\tau^2}\Big\}\Big],$$

and

$$\pi_{u2}(\boldsymbol{v},\sigma^2,\tau^2,\boldsymbol{\beta}) = (\sigma^2)^{-(N+1)/2}(\tau^2)^{-z/2}\pi_*(\tau^2)\prod_{i=1}^{m}\delta_0(v_i)^{1-u_i} \times\prod_{i=1}^{m}\Big[\exp\Big\{-\frac{1}{2\sigma^2}\sum_{j=1}^{n_i}(y_{ij}-\boldsymbol{x}_{ij}^t\boldsymbol{\beta}-v_i)^2-\frac{u_i v_i^2}{2\tau^2}\Big\}\Big].$$

To show the integrability of π_{u1} and π_{u2}, we consider the case of $\boldsymbol{u}$ with $u_1 + \cdots + u_m = k$. Without loss of generality, we assume that $u_i = 1$ for $i = 1, \ldots, k$ and $u_i = 0$ for $i = k+1, \ldots, m$. Then $\pi_{u1}(\boldsymbol{v}, \sigma^2, \tau^2, \boldsymbol{\beta})$ can be rewritten as

$$\begin{aligned}
&\pi_{u1}(\boldsymbol{v}, \sigma^2, \tau^2, \boldsymbol{\beta}) \\
&\quad = (\sigma^2)^{-(N+1)/2}(\tau^2)^{-(k+1)/2} \prod_{i=1}^{k} \Big[\exp\Big\{ -\frac{1}{2\sigma^2} \sum_{j=1}^{n_i} (y_{ij} - \boldsymbol{x}_{ij}^t \boldsymbol{\beta} - v_i)^2 - \frac{v_i^2}{2\tau^2} \Big\} \Big] \\
&\qquad \times \Big[\prod_{i=k+1}^{m} \exp\Big\{ -\frac{1}{2\sigma^2} \sum_{j=1}^{n_i} (y_{ij} - \boldsymbol{x}_{ij}^t \boldsymbol{\beta})^2 \Big\} \delta_0(v_i) \Big].
\end{aligned}$$

We define an N-dimensional vector $\boldsymbol{s}(\boldsymbol{v}_*) = (\boldsymbol{s}_{(1)}(\boldsymbol{v}_*)^t, \boldsymbol{s}_{(2)}^t)^t$ as

$$\boldsymbol{s}_{(1)}(\boldsymbol{v}_*) = ((\boldsymbol{y}_1 - v_i \mathbf{1}_{n_i})^t, \ldots, (\boldsymbol{y}_k - v_k \mathbf{1}_{n_k})^t)^t$$

and $\boldsymbol{s}_{(2)} = (\boldsymbol{y}_{k+1}^t, \ldots, \boldsymbol{y}_m^t)^t$ for $\boldsymbol{v}_* = (v_1, \ldots, v_k)^t$. Then, if $N > q$, we have

$$\begin{aligned}
&\int \pi_{u1}(\boldsymbol{v}, \sigma^2, \tau^2, \boldsymbol{\beta}) d\boldsymbol{\beta} \\
&\quad \propto (\sigma^2)^{-(N-q-1)/2-1} (\tau^2)^{-(k-1)/2-1} \exp\Big\{ -\frac{\boldsymbol{s}(\boldsymbol{v}_*)^t \boldsymbol{A} \boldsymbol{s}(\boldsymbol{v}_*)}{2\sigma^2} - \frac{1}{2\tau^2} \sum_{i=1}^{k} v_i^2 \Big\},
\end{aligned} \tag{7.12}$$

where $\boldsymbol{A} = \boldsymbol{I}_N - \boldsymbol{X}(\boldsymbol{X}^t \boldsymbol{X})^{-1} \boldsymbol{X}^t$. The right-hand side is integrable with respect to σ^2 and τ^2 since $N > q + 1$ and $k \geq a > 1$, whereby we obtain

$$\int \pi_{u1}(\boldsymbol{v}, \sigma^2, \tau^2, \boldsymbol{\beta}) d\boldsymbol{\beta} d\sigma^2 d\tau^2 \propto \pi_{u1}(\boldsymbol{v}_*) \prod_{i=k+1}^{m} \delta_0(v_i),$$

where

$$\pi_{u1}(\boldsymbol{v}_*) = \{\boldsymbol{s}(\boldsymbol{v}_*)^t \boldsymbol{A} \boldsymbol{s}(\boldsymbol{v}_*)\}^{-(N-q-1)/2} (\boldsymbol{v}_*^t \boldsymbol{v}_*)^{-(k-1)/2}.$$

In what follows, we show that $\pi_{u1}(\boldsymbol{v}_*)$ is integrable. To this end, we note that

$$\int_{\mathbb{R}^k} \pi_{u1}(\boldsymbol{v}_*) d\boldsymbol{v} = \int_{\boldsymbol{v}_*^t \boldsymbol{v}_* \leq 1} \pi_{u1}(\boldsymbol{v}_*) d\boldsymbol{v} + \int_{\boldsymbol{v}_*^t \boldsymbol{v}_* \geq 1} \pi_{u1}(\boldsymbol{v}_*) d\boldsymbol{v},$$

and we evaluate the two integrals separately. For the first term, since $\boldsymbol{A}$ is idempotent and $\mathrm{rank}(\boldsymbol{A}) = N - q$ (> 0), there exists $c(\boldsymbol{y}) > 0$ such that $\boldsymbol{s}(\boldsymbol{v}_*)^t \boldsymbol{A} \boldsymbol{s}(\boldsymbol{v}_*) > c(\boldsymbol{y})$ for all $\boldsymbol{v}_*$. Then we have

$$\begin{aligned}
\int_{\boldsymbol{v}_*^t \boldsymbol{v}_* \leq 1} \pi_{u1}(\boldsymbol{v}_*) d\boldsymbol{v} &\leq c^{-(N-q-1)/2} \int_{\boldsymbol{v}_*^t \boldsymbol{v}_* \leq 1} (\boldsymbol{v}_*^t \boldsymbol{v}_*)^{-(k-1)/2} d\boldsymbol{v} \\
&= c^{-(N-q-1)/2} V(S^k) \int_0^1 (r^2)^{-(k-1)/2} (r^2)^{(k-1)/2} dr < \infty,
\end{aligned}$$

where $V(S^k)$ is the volume of the unit sphere in $\mathbb{R}^k$. For the second term, it follows that

$$\int_{\boldsymbol{v}_*^t \boldsymbol{v}_* \geq 1} \pi_{u1}(\boldsymbol{v}_*) d\boldsymbol{v} = \int_{\boldsymbol{v}_*^t \boldsymbol{v}_* \geq 1} \{\boldsymbol{s}(\boldsymbol{v}_*)^t \boldsymbol{A} \boldsymbol{s}(\boldsymbol{v}_*)\}^{-(N-q-1)/2} (\boldsymbol{v}_*^t \boldsymbol{v}_*)^{-(k-1)/2} d\boldsymbol{v}.$$

Since $\boldsymbol{s}(\boldsymbol{v}_*)^t\boldsymbol{A}\boldsymbol{s}(\boldsymbol{v}_*)$ is a quadratic function of $\boldsymbol{v}_*$, the integral is finite provided that $N > q+2$. For the integrability of π_{u_2}, we carry out integration with respect to $\boldsymbol{\beta}, \sigma^2$ and τ^2 to get

$$\int \pi_{u2}(\boldsymbol{v}, \sigma^2, \tau^2, \boldsymbol{\beta}) d\boldsymbol{\beta} d\sigma^2 d\tau^2 \propto \pi_{u2}(\boldsymbol{v}_*) \prod_{i=k+1}^{m} \delta_0(v_i).$$

Since for $N > q + 1$,

$$\begin{aligned}\pi_{u2}(\boldsymbol{v}_*) =& \{\boldsymbol{s}(\boldsymbol{v}_*)^t\boldsymbol{A}\boldsymbol{s}(\boldsymbol{v}_*)\}^{-(N-q-1)/2}(\boldsymbol{v}_*^t\boldsymbol{v}_* + 2b_2)^{-k/2-b_1} \\ \leq& c^{-(N-q-1)/2}(2b_2)^{-k/2-b_1},\end{aligned}$$

it follows that $\pi_{u2}(\boldsymbol{v}_*)$ is integrable so long as $N > q + 1$. Thus the proof of part (a) is established.

For the proof of part (b), it is sufficient to show that the posterior second moments are finite. Since the statement for p is clear, we establish the result for $\boldsymbol{\beta}, \sigma^2$ and τ^2. As in the proof of part (a), we consider the three cases where $z > a$, $0 < z \leq a$ and $z = 0$. By replacing $N + 1$ in expressions of π_{u1}, π_{u2} and π_{u3} with $N + 5$, it follows that $\mathrm{E}\{(\sigma^2)^2|D\} < \infty$ when $N > q + 6$.

For $\mathrm{E}(\boldsymbol{\beta}\boldsymbol{\beta}^t|D)$, we first note that

$$\begin{aligned}&\int_{\mathbb{R}^q} \boldsymbol{\beta}\boldsymbol{\beta}^t \exp\Big[- \frac{\{\boldsymbol{s}(\boldsymbol{v}_*) - \boldsymbol{X}\boldsymbol{\beta}\}^t\{\boldsymbol{s}(\boldsymbol{v}_*) - \boldsymbol{X}\boldsymbol{\beta}\}}{2\sigma^2}\Big] d\boldsymbol{\beta} \\ &=(\sigma^2)^{q/2}|\boldsymbol{X}^t\boldsymbol{X}|^{-1/2} \exp\Big\{ - \frac{\boldsymbol{s}(\boldsymbol{v}_*)^t\boldsymbol{A}\boldsymbol{s}(\boldsymbol{v}_*)}{2\sigma^2}\Big\}(\boldsymbol{X}^t\boldsymbol{X})^{-1} \\ &\quad \times \{\sigma^2\boldsymbol{I}_q + \boldsymbol{X}^t\boldsymbol{s}(\boldsymbol{v}_*)\boldsymbol{s}(\boldsymbol{v}_*)^t\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\} \\ &=(\sigma^2)^{(q+2)/2}h(\boldsymbol{X}, \boldsymbol{s}(\boldsymbol{v}_*), \sigma^2) \\ &\quad + (\sigma^2)^{q/2}h(\boldsymbol{X}, \boldsymbol{s}(\boldsymbol{v}_*), \sigma^2)\boldsymbol{X}^t\boldsymbol{s}(\boldsymbol{v}_*)\boldsymbol{s}(\boldsymbol{v}_*)^t\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1},\end{aligned}$$

for

$$h(\boldsymbol{X}, \boldsymbol{s}(\boldsymbol{v}_*), \sigma^2) = |\boldsymbol{X}^t\boldsymbol{X}|^{-1/2} \exp\Big\{ - \frac{1}{2\sigma^2}\boldsymbol{s}(\boldsymbol{v}_*)^t\boldsymbol{A}\boldsymbol{s}(\boldsymbol{v}_*)\Big\}(\boldsymbol{X}^t\boldsymbol{X})^{-1}.$$

Then it follows that

$$\begin{aligned}&\int \boldsymbol{\beta}\boldsymbol{\beta}^t\pi_{u1}(\boldsymbol{v}, \boldsymbol{\beta}, \sigma^2, \tau^2) d\boldsymbol{v} d\boldsymbol{\beta} d\sigma^2 d\tau^2 \\ &\qquad \propto \boldsymbol{I}_q \int_{\mathbb{R}^k} \{\boldsymbol{s}(\boldsymbol{v}_*)^t\boldsymbol{A}\boldsymbol{s}(\boldsymbol{v}_*)\}^{-(N-q-3)/2}(\boldsymbol{v}_*^t\boldsymbol{v}_*)^{-(k-1)/2} d\boldsymbol{v} \\ &\qquad\quad + \int_{\mathbb{R}^k} \boldsymbol{X}^t\boldsymbol{s}(\boldsymbol{v}_*)\boldsymbol{s}(\boldsymbol{v}_*)^t\boldsymbol{X}\pi_{u1}(\boldsymbol{v}_*) d\boldsymbol{v}.\end{aligned}$$

Since $\boldsymbol{v}_*\boldsymbol{v}_*^t \leq (\boldsymbol{v}_*^t\boldsymbol{v}_*)\boldsymbol{I}_q$, the second term is finite if $k > 5$ for all $k \geq a$, namely $a > 5$. The first term is also finite whenever $N > q + 4$.

For the cases $0 < z \leq a$ and $z = 0$, we can similarly show that

$$\int \boldsymbol{\beta}\boldsymbol{\beta}^t\pi_{u2}(\boldsymbol{v}, \boldsymbol{\beta}, \sigma^2, \tau^2) d\boldsymbol{v} d\boldsymbol{\beta} d\sigma^2 d\tau^2 < \infty$$

and

$$\int \boldsymbol{\beta}\boldsymbol{\beta}^t \pi_{u3}(\boldsymbol{v}, \boldsymbol{\beta}, \sigma^2, \tau^2) d\boldsymbol{v} d\boldsymbol{\beta} d\sigma^2 d\tau^2 < \infty$$

under the conditions given in Theorem 7.1.

Finally, for $\mathrm{E}\{(\tau^2)^2|D\}$, it follows that

$$\begin{aligned}
&\int (\tau^2)^2 \pi_{u1}(\boldsymbol{v}, \boldsymbol{\beta}, \sigma^2, \tau^2) d\boldsymbol{v} d\boldsymbol{\beta} d\sigma^2 d\tau^2 \\
&\propto \int (\sigma^2)^{-(N-p-1)/2-1} (\tau^2)^{-(k-5)/2-1} \\
&\quad \times \exp\Big\{ -\frac{\boldsymbol{s}(\boldsymbol{v}_*)^t \boldsymbol{A} \boldsymbol{s}(\boldsymbol{v}_*)}{2\sigma^2} - \frac{1}{2\tau^2} \sum_{i=1}^{k} v_i^2 \Big\} d\boldsymbol{v} d\tau^2 d\sigma^2 \\
&\propto \int \{\boldsymbol{s}(\boldsymbol{v}_*)^t \boldsymbol{A} \boldsymbol{s}(\boldsymbol{v}_*)\}^{-(N-p-1)/2} (\boldsymbol{v}_*^t \boldsymbol{v}_*)^{-(k-5)/2} d\boldsymbol{v}
\end{aligned}$$

which is finite provided that $k > 5$ for all $k \geq a$, namely $a > 5$. In the cases $0 < z \leq a$ and $z = 0$, it is integrable if

$$\int_0^\infty (\tau^2)^2 \pi_*(\tau^2) d\tau^2 < \infty,$$

which can be established since $b_1 > 3$. Thus the proof of part (b) is complete.

Chapter 8

Empirical Uncertain Bayes Methods

8.1 Introduction

While Datta and Mandal (2015) proposed the uncertain random effects in the traditional Fay-Herriot model, the model is restrictive and it cannot be used for count or binary data. Hence, in this chapter, we focus on mixed models based on the natural exponential family as introduced in Section 4.3. To implement the idea of uncertain random effects in this context, we rewrite the uncertain random effect model as the hierarchical form:

$$y_i|\theta_i \sim N(0, D_i), \quad \theta_i|(s_i = 1) \sim N(\boldsymbol{x}_i^t\boldsymbol{\beta}, A), \quad \theta_i|(s_i = 0) = \boldsymbol{x}_i^t\boldsymbol{\beta}.$$

This means that one of the two distributions $N(\boldsymbol{x}_i^t\boldsymbol{\beta}, A)$ and the one point distribution on $\boldsymbol{x}_i^t\boldsymbol{\beta}$ is randomly selected for the prior distribution of θ_i in each area. In this paper, we naturally extend the idea to the NEF-QVF family. Since there exist the conjugate priors for the natural parameter θ_i, we introduce the uncertain prior for θ_i, a mixture distribution of the conjugate prior and the one-point distribution on the synthetic mean.

For estimating area means under the model, we here develop an empirical Bayes (EB) approach while Datta and Mandal (2015) considered a hierarchical Bayes (HB) approach. In the normal case as considered in Datta and Mandal (2015), a full Bayesian approach is relatively attractive since all the full conditional distribution of the model parameters as well as the random effects have familiar forms, so that we can efficiently sample from the posterior distribution using a Gibbs sampling. However, in the non-normal case, the posterior distribution of the model parameters are not necessarily in familiar forms, so that we need to rely on an inefficient sampling algorithm such as a Metropolis-Hastings algorithm. Moreover, the HB approach requires checking prior sensitivity and monitoring the convergence of the MCMC algorithm. As suggested in Datta and Mandal (2015), the use of non-informative (improper) enables us to avoid subjective specification of priors, but the posterior propriety is not straightforward under non-normal cases. On the other hand, the empirical Bayes approach can enjoy easily computing point estimates of model parameters and Bayes estimator without requiring prior distributions. Since one of the greatest purposes in small area estimation is point estimation, the EB approach is more attractive in this case.

Owing to the conjugacy of the prior distribution, we can easily establish the Expectation-Maximization (EM) algorithm for maximizing the marginal likelihood function to get the estimates of model parameters. Using the estimator, we derive the the empirical uncertain

Bayes (EUB) estimator of the area mean. For calibration of uncertainty of the EUB estimator, we consider the conditional mean squared error (CMSE) and derive the second-order unbiased estimator motivated from the work of Booth and Hobert (1998), Datta et al. (2011) and Sugasawa and Kubokawa (2016a). As typical examples, we handle three models, namely the Fay-Herriot model for continuous data, the Poisson-gamma and binomial-beta models for count data. It is shown that the shrinkage property pointed out in Datta and Mandal (2015) in the Fay-Herriot model still holds in both the Poisson-gamma and binomial-beta models. That is, the shrinkage coefficient in the EUB estimator decreases as the y_i gets close to the synthetic mean.

This chapter is organized as follows. In Section 8.2, we provide the detailed description of the proposed model, the EM algorithm for parameter estimation and three examples. In Section 8.3, we derive the second order unbiased estimator of CMSE for risk evaluation of the EUB estimator. Simulation studies and empirical applications are given in 8.4 and 8.5, respectively.

8.2 Empirical Uncertain Bayes Methods

8.2.1 Model setup and uncertain Bayes estimator

Let $y_1, \ldots, y_m$ be mutually independent random variables where the conditional distribution of y_i given θ_i belongs to the the following natural exponential family:

$$y_i|\theta_i \sim f(y_i|\theta_i) = \exp\{n_i(\theta_i y_i - \psi(\theta_i)) + c(y_i, n_i)\}, \tag{8.1}$$

where n_i is a known scalar parameter and is not necessarily corresponding to the sample size in the ith area. As the prior distribution of θ_i, we set the uncertain random structure treated in Datta and Mandal (2015). Let $s_1, \ldots, s_m$ be mutually independent and identical random variables distributed as

$$P(s_i = 1) = p, \quad P(s_i = 0) = 1 - p.$$

The prior distribution of θ_i is given by

$$\theta_i|(s_i = 1) \sim \pi(\theta_i) = \exp\left\{\nu(m_i\theta_i - \psi(\theta_i)) + C(\nu, m_i)\right\}, \quad \theta_i|(s_i = 0) = (\psi')^{-1}(m_i), \tag{8.2}$$

where ν is an unknown scalar hyperparameter, $C(\nu, m_i)$ is the normalizing constant and $\psi'(t) = d\psi(t)/dt$. In our settings, we consider the canonical link

$$m_i = \psi'(\boldsymbol{x}_i^t\boldsymbol{\beta}),$$

where $\boldsymbol{x}_i$ is a $q \times 1$ vector of explanatory variables, $\boldsymbol{\beta}$ is a $q \times 1$ unknown common vector of regression coefficients. The function $f(y_i|\theta_i)$ is the regular one-parameter exponential family and the function $\pi(\theta_i)$ is the conjugate prior distribution. Then the unknown parameter in two-stage model (8.1) and (8.2) are $\boldsymbol{\phi} = (\boldsymbol{\beta}^t, \nu, p)^t$. The quantity of interest in this paper is the conditional expectation of y_i given θ_i, defined as

$$\mu_i = \mathrm{E}[y_i|\theta_i] = \psi'(\theta_i),$$

noting that $\mu_i|(s_i = 0) = m_i$ from (8.2). For $\psi''(t) = d^2\psi(t)/dt^2$, we assume that $\psi''(\theta_i) = Q(\mu_i)$, namely,

$$\mathrm{Var}(y_i|\theta_i) = \frac{\psi''(\theta_i)}{n_i} = \frac{Q(\mu_i)}{n_i},$$

with $Q(x) = v_0 + v_1 x + v_2 x^2$ for known constants v_0, v_1 and v_2 which are not simultaneously zero. This means that the conditional variance $\mathrm{Var}(y_i|\theta_i)$ is a quadratic function of the conditional expectation $\mathrm{E}[y_i|\theta_i]$. Similarly, the mean and variance of the prior distribution given $s_i = 1$ are

$$\mathrm{E}[\mu_i|s_i = 1] = m_i, \quad \mathrm{Var}(\mu_i|s_i = 1) = \frac{Q(m_i)}{\nu - v_2}.$$

The joint density (or mass) function of (y_i, θ_i, s_i) is

$$g(y_i, \theta_i, s_i = 1) = f(y_i|\theta_i)\pi(\theta_i), \quad g(y_i, \theta_i, s_i = 0) = \delta_{\theta_i}((\psi')^{-1}(m_i))f(y_i|\theta_i),$$

where $\delta_{\theta_i}(a)$ denotes the point mass at $\theta_i = a$. Then the joint distribution of (y_i, θ_i) and the marginal distribution of y_i are both mixtures of two distributions:

$$\begin{aligned} g(y_i, \theta_i) &= pf(y_i|\theta_i)\pi(\theta_i) + (1-p)\delta_{\theta_i}((\psi')^{-1}(m_i))f(y_i|\theta_i), \\ f(y_i; \boldsymbol{\phi}) &= pf_1(y_i; \boldsymbol{\phi}) + (1-p)f_2(y_i; \boldsymbol{\phi}), \end{aligned}$$

where

$$f_1(y_i; \boldsymbol{\phi}) = \int f(y_i|\theta_i)\pi(\theta_i)d\theta_i, \quad f_2(y_i; \boldsymbol{\phi}) = f(y_i|\theta_i = (\psi')^{-1}(m_i)). \tag{8.3}$$

Since $\pi(\theta_i)$ is the conjugate prior of θ_i, the marginal distribution $f_1(y_i; \boldsymbol{\phi})$ and the conditional distribution $\pi(\theta_i|y_i, s_i = 1; \boldsymbol{\phi})$ can be obtain in the closed forms:

$$\begin{aligned} \pi(\theta_i|y_i, s_i = 1; \boldsymbol{\phi}) &= \exp\big\{(n_i + \nu)(\eta_i\theta_i - \psi(\theta_i))\big\}C(n_i + \nu, \eta_i), \\ f_1(y_i; \boldsymbol{\phi}) &= \frac{C(\nu, m_i)}{C(n_i + \nu, \eta_i)} \exp\big\{c(y_i, n_i)\big\}, \end{aligned}$$

where

$$\eta_i \equiv \eta_i(y_i; \boldsymbol{\phi}) = \frac{n_i y_i + \nu m_i}{n_i + \nu}.$$

The conditional distribution of s_i given y_i can be obtained as

$$P(s_i = 1|y_i; \boldsymbol{\phi}) = \frac{pf_1(y_i, \boldsymbol{\phi})}{f(y_i; \boldsymbol{\phi})} = \frac{p}{p + (1-p)f_2(y_i; \boldsymbol{\phi})/f_1(y_i; \boldsymbol{\phi})} = 1 - P(s_i = 0|y_i; \boldsymbol{\phi}).$$

To obtain the Bayes estimator of μ_i, we note that

$$\mathrm{E}\left[\mu_i|s_i, y_i; \boldsymbol{\phi}\right] = m_i + \frac{n_i}{\nu + n_i}(y_i - m_i)I(s_i = 1), \tag{8.4}$$

where $I(\cdot)$ is an indicator function. Hence the Bayes estimator of μ_i is

$$\widetilde{\mu}_i(y_i, \boldsymbol{\phi}) = \mathrm{E}\left[\mu_i|y_i; \boldsymbol{\phi}\right] = \mathrm{E}\left[\mathrm{E}\left(\mu_i|s_i, y_i; \boldsymbol{\phi}\right)|y_i; \boldsymbol{\phi}\right] = m_i + \frac{n_i}{\nu + n_i}(y_i - m_i)r_i(y_i, \boldsymbol{\phi}), \tag{8.5}$$

where

$$r_i(y_i, \boldsymbol{\phi}) = P(s_i = 1|y_i; \boldsymbol{\phi}) = \frac{p}{p + (1-p)f_2(y_i; \boldsymbol{\phi})/f_1(y_i; \boldsymbol{\phi})}. \tag{8.6}$$

It is observed that $r_i(y_i, \boldsymbol{\phi})$ increases in p and decreases in the ratio $f_2(y_i; \boldsymbol{\phi})/f_1(y_i; \boldsymbol{\phi})$. In what follows, we use the abbreviated notations $\widetilde{\mu}_i$ and r_i instead of $\widetilde{\mu}_i(y_i, \boldsymbol{\phi})$ and $r_i(y_i, \boldsymbol{\phi})$, respectively, when there is no confusion. It is noted that the Bayes estimator (8.5) can be expressed as

$$\widetilde{\mu}_i = y_i - \left\{1 - \frac{n_i}{\nu + n_i} r_i(y_i, \boldsymbol{\phi})\right\}(y_i - m_i),$$

which shrinks the direct estimator y_i toward the regression (or synthetic) part $m_i = \psi'(\boldsymbol{x}_i^t \boldsymbol{\beta})$, and the shrinkage function depends on y_i through r_i. On the other hand, in the classical two-stage model as used in Ghosh and Maiti (2004), the shrinkage function does not depends on the observation y_i, which is sometimes not flexible for real data analysis. It should be noted that $r_i = 1$ when $p = 1$. Thus, the suggested method includes the classical method as well as it has the shrinkage function adjusted by y_i which arises from introducing the weight parameter p. Moreover, when the prior is completely singular, namely $p = 0$, it follows $r_i = 0$ and the resulting Bayes estimator is m_i. We call the estimator (8.5) under the prior (8.2) the uncertain Bayes estimator in order to distinguish from the conventional Bayes estimator.

8.2.2 Maximum likelihood estimation using EM algorithm

Since the uncertain Bayes estimator (8.5) depends on the unknown model parameter $\boldsymbol{\phi}$, we need to estimate them for practical use. A reasonable method is the maximum likelihood (ML) estimator which maximizes the marginal distribution of y_i. Since the marginal density is the mixture of the two distributions $f(y_i; \boldsymbol{\phi}) = pf_1(y_i; \boldsymbol{\phi}) + (1-p)f_2(y_i; \boldsymbol{\phi})$, the ML estimator is the maximizer of the log-likelihood function

$$L(\boldsymbol{\phi}) = \sum_{i=1}^{m} \log\left\{pf_1(y_i; \boldsymbol{\phi}) + (1-p)f_2(y_i; \boldsymbol{\phi})\right\}. \tag{8.7}$$

To compute the ML estimate, we propose Expectation-Maximization (EM) algorithm (Dempster, Laird and Rubin, 1977) which maximizes the objective function (8.7) iteratively and indirectly. From (8.1) and (8.2), the complete log-likelihood function $L^c(\boldsymbol{\phi})$ given $(y_i, \theta_i, s_i)_{i=1,\ldots,m}$ is

$$\begin{aligned} L^c(\boldsymbol{\phi}) = &\sum_{i=1}^{m} \{n_i(\theta_i y_i - \psi(\theta_i))\} + \sum_{i=1}^{m} s_i \{\nu(m_i\theta_i - \psi(\theta_i)) + C(\nu, m_i)\} \\ &+ \sum_{i=1}^{m} \{s_i \log p + (1 - s_i)\log(1-p)\}. \end{aligned}$$

In the rth iteration, we first compute the expectation of the complete log-likelihood $\mathrm{E}^{(r)}[L^c(\boldsymbol{\phi})]$ at the E-step, where $E^{(r)}$ denotes the expectation with respect to the conditional distributions $(\theta_i, s_i)|y_i$ with hyperparameter values $\boldsymbol{\phi}^{(r)}$. Then the objective function to be maximized at the M-step in the rth iteration is

$$\begin{aligned} Q^{(r)}(\boldsymbol{\phi}) \equiv \mathrm{E}^{(r)}[L^c(\boldsymbol{\phi})] = &\sum_{i=1}^{m} r_i(y_i, \boldsymbol{\phi}^{(r)}) \left\{\nu m_i \mathrm{E}^{(r)}[\theta_i | s_i = 1] - \nu \mathrm{E}^{(r)}[\psi(\theta_i)|s_i = 1] + C(\nu, m_i)\right\} \\ &+ \sum_{i=1}^{m} \left\{r_i(y_i, \boldsymbol{\phi}^{(r)}) \log p + (1 - r_i(y_i, \boldsymbol{\phi}^{(r)}))\log(1-p)\right\}, \end{aligned}$$

which yields the updating algorithm as

$$\begin{aligned}(\boldsymbol{\beta}^{(r+1)}, \nu^{(r+1)}) &= \text{argmax}_{\boldsymbol{\beta},\nu} \sum_{i=1}^{m} r_i(y_i, \boldsymbol{\phi}^{(r)}) h_i^{(r)}(\boldsymbol{\beta}, \nu) \\ p^{(r+1)} &= \frac{1}{m} \sum_{i=1}^{m} r_i(y_i, \boldsymbol{\phi}^{(r)}),\end{aligned} \tag{8.8}$$

where $h_i^{(r)}(\boldsymbol{\beta}, \nu) = \nu m_i \mathrm{E}^{(r)}[\theta_i | s_i = 1] - \nu \mathrm{E}^{(r)}[\psi(\theta_i) | s_i = 1] + C(\nu, m_i)$. Since the prior distribution of θ_i given $s_i = 1$ is conjugate, the posterior distribution of θ_i given $s_i = 1$ belongs to the same family as the prior distribution, and we can easily generate samples from the distribution in common models as demonstrated in the subsequent section. Hence, the calculation of two expectations $\mathrm{E}^{(r)}[\theta_i | s_i = 1]$ and $\mathrm{E}^{(r)}[\psi(\theta_i) | s_i = 1]$ given in the E-step is easy to carry out. We summarize the EM algorithm in the following.

Algorithm 8.1 (EM algorithm)**.** *Iterative,*

1. *Set the initial value* $\boldsymbol{\phi}^{(0)}$ *and* $r = 0$.
2. *Compute* $\mathrm{E}^{(r)}[\theta_i | s_i = 1]$ *and* $\mathrm{E}^{(r)}[\psi(\theta_i) | s_i = 1]$ *using the current parameter value* $\boldsymbol{\phi}^{(r)}$.
3. *Update the parameter value as* $\boldsymbol{\phi}^{(r+1)}$ *based on* (8.8).
4. *If the difference between* $\boldsymbol{\phi}^{(r)}$ *and* $\boldsymbol{\phi}^{(r+1)}$ *is sufficiently small, then the estimate is given by* $\boldsymbol{\phi}^{(r+1)}$. *Otherwise, set* $r = r + 1$ *and go back to Step 2.*

Finally, substituting $\widehat{\boldsymbol{\phi}}$ into the UB estimator, we get the empirical uncertain Bayes (EUB) estimator

$$\widehat{\mu}_i \equiv \widetilde{\mu}_i(y_i, \widehat{\boldsymbol{\phi}}) = \widehat{m}_i + \frac{n_i}{\widehat{\nu} + n_i}(y_i - \widehat{m}_i) r_i(y_i, \widehat{\boldsymbol{\phi}}), \tag{8.9}$$

where $\widehat{m}_i = \psi'(\boldsymbol{x}_i^t \widehat{\boldsymbol{\beta}})$.

8.2.3 Some examples

Here we provide three typical models often used in practice and investigate properties of the UB estimators with detailed expressions of E-step and M-step in the EM algorithm.

[1] Normal-normal (Fay-Herriot) model. The Fay-Herriot model (Fay and Herriot, 1979) is an area-level model frequently used in small area estimation, given by

$$y_i | \theta_i \sim N(\theta_i, D_i), \quad \theta_i | (s_i = 1) \sim N(\boldsymbol{x}_i^t \boldsymbol{\beta}, A), \quad i = 1, \ldots, m,$$

corresponding to $n_i = D_i^{-1}, v_0 = 1, v_1 = v_2 = 0, \nu = A^{-1}$ and $\psi(\theta_i) = \theta_i^2/2$ in (8.1) and (8.2). This model was studied in Datta and Mandal (2015) in terms of Bayesian perspectives. The marginal distributions of $f_1(y_i)$ and $f_2(y_i)$ in (8.3) are given by

$$f_1(y_i; \boldsymbol{\phi}) = \frac{1}{\sqrt{2\pi(A + D_i)}} \exp\left(-\frac{(y_i - m_i)^2}{2(A + D_i)}\right), \quad f_2(y_i; \boldsymbol{\phi}) = \frac{1}{\sqrt{2\pi D_i}} \exp\left(-\frac{(y_i - m_i)^2}{2D_i}\right), \tag{8.10}$$

so that $r_i(y_i, p)$ is obtained from (8.6) as

$$r_i(y_i,p) = p\left\{p + (1-p)\sqrt{\frac{A+D_i}{D_i}}\exp\left(-\frac{A(y_i-m_i)^2}{2D_i(A+D_i)}\right)\right\}^{-1},$$

which coincides with the result given in Datta and Mandal (2015). It is clear that $r_i(y_i, p)$ takes small values when y_i is close to m_i, corresponding to the case where y_i is well explained by m_i without random effects.

Regarding the parameter estimation via the EM algorithm in the Fay-Heriot model, the objective function at the M-step is

$$Q^{(r)}(\boldsymbol{\beta}, A) = -\frac{1}{2}\sum_{i=1}^{m} r_i^{(r)}\left\{\log A + \frac{1}{A}\left(\theta_i^{(r)} - \boldsymbol{x}_i^t\boldsymbol{\beta}\right)^2\right\},$$

where $r_i^{(r)} = r_i(y_i, \boldsymbol{\beta}^{(r)}, A^{(r)}, p^{(r)})$ and $\theta_i^{(r)} = (A^{(r)}y_i + D_i\boldsymbol{x}_i^t\boldsymbol{\beta}^{(r)})/(A^{(r)} + D_i)$, so that the updating step for $\boldsymbol{\beta}$ and A is written as

$$\boldsymbol{\beta}^{(r+1)} = \left(\sum_{i=1}^{m} r_i^{(r)}\boldsymbol{x}_i\boldsymbol{x}_i^t\right)^{-1}\sum_{i=1}^{m} r_i^{(r)}\theta_i^{(r)}\boldsymbol{x}_i, \qquad A^{(r+1)} = \frac{1}{m}\sum_{i=1}^{m}\left(\theta_i^{(r)} - \boldsymbol{x}_i^t\boldsymbol{\beta}^{(r+1)}\right)^2.$$

[2] Poisson-gamma model. Let $z_1, \ldots, z_m$ be mutually independent random variables having

$$z_i|\lambda_i \sim \mathrm{Po}(n_i\lambda_i), \qquad \lambda_i|(s_i = 1) \sim \mathrm{Ga}(\nu m_i, \nu)$$

where $\lambda_1, \ldots, \lambda_m$ are mutually independent, $\mathrm{Po}(\lambda)$ denotes the Poisson distribution with mean λ, and $\mathrm{Ga}(a, b)$ denotes the gamma distribution with density $f(x) \propto x^{a-1}\exp(-bx)$. Let $y_i = z_i/n_i$ and $\log m_i = \boldsymbol{x}_i^t\boldsymbol{\beta}$ for $i = 1, \ldots, m$. Then, the notations in (8.1) and (8.2) correspond to $v_1 = 1$, $v_0 = v_2 = 0$ and $\psi(\theta_i) = \exp(\theta_i)$. The marginal distributions of $f_1(y_i)$ and $f_2(y_i)$ are given by

$$f_1(y_i;\boldsymbol{\phi}) = \frac{\Gamma(n_iy_i + \nu m_i)}{\Gamma(n_iy_i+1)\Gamma(\nu m_i)}\left(\frac{n_i}{n_i+\nu}\right)^{n_iy_i}\left(\frac{\nu}{n_i+\nu}\right)^{\nu m_i}, \quad f_2(y_i;\boldsymbol{\phi}) = \frac{(n_im_i)^{n_iy_i}}{(n_iy_i)!}\exp(-n_im_i) \tag{8.11}$$

where $\Gamma(\cdot)$ denotes a gamma function, so that $r_i(y_i, p)$ is written as

$$r_i(y_i,p) = p\left\{p + (1-p)\frac{\Gamma(\nu m_i)\exp(-n_im_i)}{\Gamma(n_iy_i+\nu m_i)}(n_i+\nu)^{n_iy_i+\nu m_i}m_i^{n_iy_i}\nu^{-\nu m_i}\right\}^{-1}. \tag{8.12}$$

Unlike the Fay-Herriot model, it is not clear when $r_i(y_i, p)$ takes small values as a function of y_i. To see this property, let $h(z_i) = (n_i+\nu)^{z_i+\nu m_i}m_i^{z_i}/\Gamma(z_i+\nu m_i)$. It is noted that $r_i(y_i, p)$ depends on $z_i(= n_iy_i)$ through $h(z_i)$. It follows that

$$\frac{h(z_i+1)}{h(z_i)} = \frac{n_im_i + \nu m_i}{z_i + \nu m_i},$$

so that we have $h(z_i) \le h(z_i+1)$ for $y_i \le m_i$ and $h(z_i) \ge h(z_i+1)$ for $y_i \ge m_i$. Then, when y_i is close to m_i, $h(z_i)$ takes a large value, which results in a small value of $r_i(y_i, p)$. This observation is similar to the case of the Fay-Herriot model.

The objective function at the M-step in the EM algorithm can be expressed as

$$Q^{(r)}(\boldsymbol{\beta},\nu)=\sum_{i=1}^{m} r_i(y_i,\boldsymbol{\beta}^{(r)},\nu^{(r)},p^{(r)})\Big\{\nu m_i\log\nu-\log\Gamma(\nu m_i) \\ +\nu m_i\int_0^\infty(\log t)f_\Gamma(t;n_iy_i+\nu^{(r)}m_i^{(r)},n_i+\nu^{(r)})dt-\nu\frac{n_iy_i+\nu^{(r)}m_i^{(r)}}{n_i+\nu^{(r)}}\Big\},$$

where $f_\Gamma(\cdot;a,b)$ denotes the density function of $\mathrm{Ga}(a,b)$. It should be noted that the integral given in the objective function can be easily calculated by generating samples from $\mathrm{Ga}(n_iy_i+\nu^{(r)}m_i^{(r)},n_i+\nu^{(r)})$.

[3] Binomial-beta model. Let $z_1,\ldots,z_m$ be mutually independent random variables having

$$z_i|p_i\sim\mathrm{Bin}(n_i,p_i),\quad p_i|(s_i=1)\sim\mathrm{Beta}(\nu m_i,\nu(1-m_i)),$$

where $p_1,\ldots,p_m$ are mutually independent, $\mathrm{Bin}(n,p)$ denotes the binomial distribution and $\mathrm{Beta}(a,b)$ denotes the beta distribution with density $f(x)\propto x^{a-1}(1-x)^{b-1}$. Let $y_i=z_i/n_i$ and $m_i=\exp(\boldsymbol{x}_i^t\boldsymbol{\beta})/(1+\exp(\boldsymbol{x}_i^t\boldsymbol{\beta}))$ for $i=1,\ldots,m$. Then the notations in (8.1) and (8.2) correspond to $v_0=0,\ v_1=1$ and $v_2=-1,\ \mu_i=p_i=\exp(\theta_i)/(1+\exp(\theta_i))$ and $\psi(\theta_i)=\log(1+\exp(\theta_i))$. The marginal distributions of $f_1(y_i)$ and $f_2(y_i)$ are

$$f_1(y_i;\boldsymbol{\phi})=\binom{n_i}{n_iy_i}\frac{B(\nu m_i+n_iy_i,n_i(1-y_i)+\nu(1-m_i))}{B(\nu m_i,\nu(1-m_i))},\quad f_2(y_i;\boldsymbol{\phi})=\binom{n_i}{n_iy_i}m_i^{n_iy_i}(1-m_i)^{n_i(1-y_i)},$$

where $B(\cdot,\cdot)$ denotes a beta function, so that $r_i(y_i,p)$ is written as

$$r_i(y_i,p)=p\left\{p+(1-p)\frac{B(\nu m_i,\nu(1-m_i))}{B(\nu m_i+n_iy_i,n_i(1-y_i)+\nu(1-m_i))}m_i^{n_iy_i}(1-m_i)^{n_i(1-y_i)}\right\}^{-1}.$$

Using the same arguments as in the Poisson-gamma model, we consider the function $h(z_i)=m_i^{z_i}(1-m_i)^{n_i-z_i}/B(\nu m_i+z_i,n_i-z_i+\nu(1-m_i))$. Then the straightforward calculation shows that

$$\frac{h(z_i+1)}{h(z_i)}=\frac{m_i\{n_i-z_i-1+\nu(1-m_i)\}}{(1-m_i)(\nu m_i+z_i)},$$

whereby $h(z_i)\le h(z_i+1)$ for $y_i\le m_i(1-n_i^{-1})$ and $h(z_i)\ge h(z_i+1)$ for $y_i\ge m_i(1-n_i^{-1})$. Thus, when y_i is close to m_i, $h(z_i)$ takes a large value, which results in a small value of $r_i(y_i,p)$.

The objective function at the M-step in the EM algorithm is expressed as

$$Q^{(r)}(\boldsymbol{\beta},\nu)=\sum_{i=1}^{m} r_i(y_i,\boldsymbol{\beta}^{(r)},\nu^{(r)},p^{(r)})\Big\{\nu m_i\int_0^1(\log t)f_B(t;a_i^{(r)},b_i^{(r)})dt \\ +\nu(1-m_i)\int_0^1\log(1-t)f_B(t;a_i^{(r)},b_i^{(r)})dt-\log B(\nu m_i,\nu(1-m_i))\Big\},$$

where $a_i^{(r)}=n_iy_i+\nu^{(r)}m_i^{(r)}$, $b_i^{(r)}=n_i(1-y_i)+\nu^{(r)}(1-m_i^{(r)})$ and $f_B(\cdot;a,b)$ denotes the density function of the beta distribution $\mathrm{Beta}(a,b)$. The two integrals given in the above formula can be easily computed by generating samples from $\mathrm{Beta}(a_i^{(r)},b_i^{(r)})$.

In Figure 8.1, we draw the shrinkage function $r_i(y_i, p)$ as a function of y_i for the three models, where $n_i = 10$, $\nu = 10$, $m_i = \psi'(\beta)$ at $\beta = 0$, and the solid, dashed and dotted lines correspond to the three values $p = 0.2$, 0.5 and 0.8, respectively. It is observed from Figure 8.1 that the shrinkage function in all the models are actually minimized at $y_i = m_i$ as discussed so far, and converges to 1 as y_i goes away from m_i. Especially, it is interesting to point out that in the Poisson-gamma model, the shrinkage ratio is not symmetric around $y_i = m_i$, while the other two models are symmetric around $y_i = m_i$.

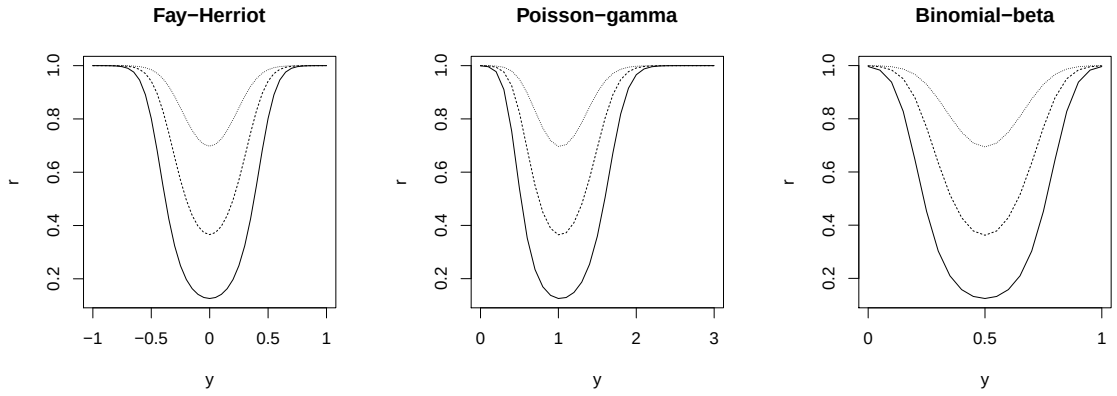

Figure 8.1: Shrinkage function $r_i(y_i, p)$ in the thee models with $p = 0.2$ (solid), 0.5 (dashed), and 0.8 (dotted).

8.3 Risk Evaluation of the EUB Estimator

8.3.1 Conditional MSE of the EUB estimator

In practice, the risk evaluation of the resulting estimator is an important issue in small area estimation. The unconditional mean squared error (MSE) is often used, but it is not suitable in this context, because researchers are interested in the risk of the area-specific risk in predicting μ_i under given y_i. This philosophy was originally proposed by Booth and Hobert (1998), and they suggested using the conditional MSE (CMSE) instead of the classical unconditional MSE in the context of mixed model prediction. Since then, the CMSE has been studied in the literature of small area estimation, including Datta et al. (2011a) and Sugasawa and Kubokawa (2016). The CMSE of the EUB estimator is defined as

$$\mathrm{CM}_i(y_i, \boldsymbol{\phi}) = \mathrm{E}\left[(\widehat{\mu}_i - \mu_i)^2 | y_i; \boldsymbol{\phi}\right],$$

noting that the expectation is taken with respect to the conditional distribution $Y_{(-i)}|y_i$ with $Y_{(-i)} = \{y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_m\}$. Because $\widetilde{\mu}_i$ is the conditional expectation, the CMSE can be decomposed as

$$\mathrm{CM}_i = \mathrm{Var}(\mu_i | y_i; \boldsymbol{\phi}) + \mathrm{E}\left[(\widehat{\mu}_i - \widetilde{\mu}_i)^2 | y_i; \boldsymbol{\phi}\right]. \tag{8.13}$$

We shall evaluate the two terms in the right hand side of (8.13).

Concerning the first term in (8.13), owing to the quadratic variance structure of the assumed model, we have

$$\mathrm{Var}(\mu_i|y_i;\boldsymbol{\phi}) = \mathrm{E}\left[\mathrm{Var}(\mu_i|s_i,y_i;\boldsymbol{\phi})|y_i;\boldsymbol{\phi}\right] + \mathrm{Var}\left(\mathrm{E}[\mu_i|s_i,y_i;\boldsymbol{\phi}]|y_i;\boldsymbol{\phi}\right).$$

In the case of $s_i = 1$, we have $\mathrm{Var}(\mu_i|s_i=1,y_i;\boldsymbol{\phi}) = Q(\eta_i)/(n_i+\nu-v_2)$ for $\eta_i = \mathrm{E}[\mu_i|s_i = 1,y_i;\boldsymbol{\phi}] = (n_iy_i+\nu m_i)/(n_i+\nu)$. Thus,

$$\mathrm{Var}(\mu_i|s_i,y_i;\boldsymbol{\phi}) = \frac{Q(\eta_i)}{n_i+\nu-v_2} I(s_i=1).$$

From (8.4), it follows that

$$\mathrm{Var}(\mu_i|y_i;\boldsymbol{\phi}) = \frac{Q(\eta_i)}{n_i+\nu-v_2}P(s_i=1|y_i;\boldsymbol{\phi}) + \mathrm{Var}\Big(m_i + \frac{n_i}{\nu+n_i}(y_i-m_i)I(s_i=1)|y_i;\boldsymbol{\phi}\Big). \quad (8.14)$$

Here it is observed that

$$\begin{aligned}\mathrm{Var}\Big(m_i + \frac{n_i}{\nu+n_i}(y_i-m_i)I(s_i=1)|y_i;\boldsymbol{\phi}\Big) &= \Big(\frac{n_i}{\nu+n_i}\Big)^2(y_i-m_i)^2 E\Big[I(s_i=1) - 2r_iI(s_i=1) + r_i^2|y_i;\boldsymbol{\phi}\Big] \\ &= \Big(\frac{n_i}{\nu+n_i}\Big)^2(y_i-m_i)^2 r_i(1-r_i),\end{aligned}$$

where r_i is given in (8.6). We thus get

$$R_{1i}(y_i,\boldsymbol{\phi}) \equiv \mathrm{Var}(\mu_i|y_i;\boldsymbol{\phi}) = \frac{n_i^2}{(\nu+n_i)^2}(y_i-m_i)^2 r_i(1-r_i) + \frac{r_iQ(\eta_i)}{n_i+\nu-v_2}, \quad (8.15)$$

which is of order $O_p(1)$.

Concerning the second term $\mathrm{E}\left[(\widehat{\mu}_i-\widetilde{\mu}_i)^2|y_i\right]$ in (8.13), we approximate it up to second order. For notational simplicity, let $\boldsymbol{\phi} = (\phi_1,\ldots,\phi_q,\phi_{q+1},\phi_{q+2})^t$ for $(\phi_1,\ldots,\phi_q)^t = \boldsymbol{\beta}$, $\phi_{q+1} = \nu$ and $\phi_{q+2} = p$. Let $\widehat{\boldsymbol{\phi}} = (\widehat{\phi}_1,\ldots,\widehat{\phi}_q,\widehat{\phi}_{q+1},\widehat{\phi}_{q+2})^t$ be the ML estimator of $\boldsymbol{\phi}$, where $(\widehat{\phi}_1,\ldots,\widehat{\phi}_q)^t = \widehat{\boldsymbol{\beta}}$, $\widehat{\phi}_{q+1} = \widehat{\nu}$ and $\widehat{\phi}_{q+2} = \widehat{p}$. The asymptotic variance and bias of $\widehat{\boldsymbol{\phi}}$ are, respectively, written as

$$\boldsymbol{\Omega} \equiv \mathrm{E}\left\{(\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi})(\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi})^t\right\}, \quad \mathrm{B} \equiv \mathrm{E}\left(\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi}\right).$$

It is noted that $\boldsymbol{\Omega}$ and B are of order $O(m^{-1})$. Assume the following regularity conditions.

Assumption 8.1.

(i) *There exist* $\underline{n},\overline{n} > 0$ *such that* $\underline{n} \le n_i \le \overline{n}$ *for all* $i = 1,\ldots,m$.

(ii) *The true value of the parameter* $\boldsymbol{\phi}$ *is in the interior of* Φ, *where* Φ *is the parameter space.*

(iii) *The densities* $f_a(y_i;\boldsymbol{\phi})$ *for* $a = 1,2$ *are three times continuously differentiable and satisfies for* $j,\ell,k = 1,\ldots,q+2$,

$$|f_{a(\phi_j)}(y_i;\boldsymbol{\phi})| + |f_{a(\phi_j\phi_\ell)}(y_i;\boldsymbol{\phi})| + |f_{a(\phi_j\phi_\ell\phi_k)}(y_i;\boldsymbol{\phi})| \le C(y_i,\boldsymbol{\phi}),$$

for fixed $\boldsymbol{\phi}$ *and* $\mathrm{E}[|C(y_i,\boldsymbol{\phi})|^{4+\delta}] < \infty$ *for some* $\delta > 0$.

The assumption (i) is a standard one in this context. For example, in the Fay-Heriot model described in Section 8.2.3, the assumption corresponds to $\underline{D} \leq D_i \leq \overline{D},\ i = 1, \ldots, m$ for some $\underline{D}$ and $\overline{D}$, which is usually assumed in the context of small area estimation (e.g. Datta et al., 2005). The assumptions (ii) and (iii) are required for deriving the asymptotic properties of the ML estimator of $\boldsymbol{\phi}$ as provided in Lemma 8.1. It should be noted that the typical three models described in Section 8.2.3 satisfy the assumption (iii), which can be demonstrated in Section 8.6.

Since $y_1, \ldots, y_m$ are mutually independent, from Theorem 1 in Lohr and Rao (2009), we can get the following lemma about asymptotic properties of estimators.

Lemma 8.1. *For the ML estimator $\widehat{\boldsymbol{\phi}}$, under Assumption 1, it holds that $\sqrt{m}(\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi}) = O_p(1)$, $\mathrm{E}\left\{(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi})(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi})^t | y_i\right\} = \boldsymbol{\Omega} + o_p(m^{-1})$ and*

$$\mathrm{E}(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}|y_i) = \mathrm{B} - \boldsymbol{\Omega} L_{i(\boldsymbol{\phi})}(y_i, \boldsymbol{\phi}) + o_p(m^{-1}),$$

where $L_{i(\boldsymbol{\phi})}(y_i, \boldsymbol{\phi}) = \partial L_i(y_i, \boldsymbol{\phi})/\partial \boldsymbol{\phi}$ for $L_i(y_i, \boldsymbol{\phi}) = \log\{p f_1(y_i; \boldsymbol{\phi}) + (1-p) f_2(y_i; \boldsymbol{\phi})\}$.

Note that the conditional asymptotic variance of $\widehat{\boldsymbol{\phi}}$ does not depend on y_i, while the conditional asymptotic bias depends on y_i. Using Lemma 8.1, we can evaluate the second term as

$$\begin{aligned}\mathrm{E}\left[(\widehat{\mu}_i - \widetilde{\mu}_i)^2 | y_i\right] &= \mathrm{E}\left[\left\{\widetilde{\mu}_{i(\boldsymbol{\phi})}^t (\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi})\right\}^2 \Big| y_i\right] + o_p(m^{-1}) \\ &= \mathrm{tr}\left(\boldsymbol{\Omega} \widetilde{\mu}_{i(\boldsymbol{\phi})} \widetilde{\mu}_{i(\boldsymbol{\phi})}^t\right) + o_p(m^{-1}),\end{aligned}$$

where $\widetilde{\mu}_{i(\boldsymbol{\phi})} = \partial \widetilde{\mu}_i / \partial \boldsymbol{\phi}$. Let

$$R_{2i}(y_i, \boldsymbol{\phi}) \equiv \mathrm{tr}\left(\boldsymbol{\Omega} \widetilde{\mu}_{i(\boldsymbol{\phi})} \widetilde{\mu}_{i(\boldsymbol{\phi})}^t\right). \tag{8.16}$$

Since $R_{2i}(y_i, \boldsymbol{\phi}) = O_p(m^{-1})$, we obtain the following theorem.

Theorem 8.1. *Let $\mathrm{CM}_i^*(y_i, \boldsymbol{\phi}) = R_{1i}(y_i, \boldsymbol{\phi}) + R_{2i}(y_i, \boldsymbol{\phi})$ for R_{1i} and R_{2i} given in* (8.15) *and* (8.16)*, respectively. Under Assumption 1, we have*

$$\mathrm{CM}_i(y_i, \boldsymbol{\phi}) = \mathrm{CM}_i^*(y_i, \boldsymbol{\phi}) + o_p(m^{-1}).$$

8.3.2 Second-order unbiased estimator of CMSE

The approximated CMSE given in Theorem 8.1 depends on the unknown parameter $\boldsymbol{\phi}$, so that it is not feasible in practice. Here we provide a second-order unbiased estimator of the CMSE. In what follows, we use the abbreviated notations R_{1i} and R_{2i} instead of $R_{1i}(y_i, \boldsymbol{\phi})$ and $R_{2i}(y_i, \boldsymbol{\phi})$, respectively, without any confusion.

Since $R_{2i} = O_p(m^{-1})$, we estimate it by the plug-in estimator $R_{2i}(y_i, \widehat{\boldsymbol{\phi}})$ with second-order unbiasedness, that is, $\mathrm{E}[R_{2i}(y_i, \widehat{\boldsymbol{\phi}}) - R_{2i}(y_i, \boldsymbol{\phi})|y_i] = o_p(m^{-1})$. On the other hand, the plug-in estimator $R_{1i}(y_i, \widehat{\boldsymbol{\phi}})$ has a second-order bias, namely $\mathrm{E}[R_{1i}(y_i, \widehat{\boldsymbol{\phi}}) - R_{1i}(y_i, \boldsymbol{\phi})|y_i] = O_p(m^{-1})$,

because $R_{1i} = O_p(1)$. To achieve the second-order accuracy, we correct the second-order bias of the estimator $R_{1i}(y_i, \widehat{\boldsymbol{\phi}})$. Using the Taylor series expansion, we have

$$R_{1i}(y_i, \widehat{\boldsymbol{\phi}}) = R_{1i} + R^t_{1i(\boldsymbol{\phi})}(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}) + \frac{1}{2}(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi})^t R_{1i(\boldsymbol{\phi}\boldsymbol{\phi})}(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}) + o_p(m^{-1}),$$

where $R_{1i(\boldsymbol{\phi})} = \partial R_{1i}/\partial \boldsymbol{\phi}$ and $R_{1i(\boldsymbol{\phi}\boldsymbol{\phi})} = \partial^2 R_{1i}/\partial \boldsymbol{\phi}\partial \boldsymbol{\phi}^t$. From Lemma 8.1, it is seen that the second-order bias in $R_{1i}(y_i, \widehat{\boldsymbol{\phi}})$ is

$$\begin{aligned} b_i(y_i, \boldsymbol{\phi}) \equiv & R^t_{1i(\boldsymbol{\phi})} E\left(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi} | y_i\right) + \frac{1}{2}\mathrm{tr}\left(R_{1i(\boldsymbol{\phi}\boldsymbol{\phi})} \mathrm{E}\left[(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi})(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi})^t | y_i\right]\right) \\ = & R^t_{1i(\boldsymbol{\phi})}\left(\mathrm{B} - \boldsymbol{\Omega} L_{i(\boldsymbol{\phi})}\right) + \frac{1}{2}\mathrm{tr}\left(R_{1i(\boldsymbol{\phi}\boldsymbol{\phi})}\boldsymbol{\Omega}\right). \end{aligned}$$

Thus, the bias-corrected estimator of R_{1i} is given by

$$R^{BC}_{1i}(y_i, \widehat{\boldsymbol{\phi}}) = R_{1i}(y_i, \widehat{\boldsymbol{\phi}}) - b_i(y_i, \widehat{\boldsymbol{\phi}}), \tag{8.17}$$

which satisfies $\mathrm{E}[R^{BC}_{1i}(y_i, \widehat{\boldsymbol{\phi}}) - R_{1i}(y_i, \boldsymbol{\phi})|y_i] = o_p(m^{-1})$.

Theorem 8.2. *Let* $\widehat{\mathrm{CM}}_i = R^{BC}_{1i}(y_i, \widehat{\boldsymbol{\phi}}) + R_{2i}(y_i, \widehat{\boldsymbol{\phi}})$, *where* $R^{BC}_{1i}(y_i, \widehat{\boldsymbol{\phi}})$ *is given in* (8.17). *Then, under Assumption 1, we have*

$$\mathrm{E}\left[\widehat{\mathrm{CM}}_i - \mathrm{CM}_i | y_i\right] = o_p(m^{-1}).$$

To calculate the $\widehat{\mathrm{CM}}_i$, we compute the estimates of $\boldsymbol{\Omega}$ and B using the parametric bootstrap method. Let $\widehat{\boldsymbol{\Omega}}$ and $\widehat{B}$ be bootstrap estimators of $\boldsymbol{\Omega}$ and B, respectively. Then, we have the approximations

$$E[\widehat{\boldsymbol{\Omega}}] = \boldsymbol{\Omega} + o(m^{-1}), \quad E[\widehat{B}] = \mathrm{B} + o(m^{-1}),$$

because $\boldsymbol{\Omega} = O(m^{-1})$ and $\mathrm{B} = O(m^{-1})$. Moreover, we need to compute $f_{1(\boldsymbol{\phi})}$, $f_{2(\boldsymbol{\phi})}$, $R_{1i(\boldsymbol{\phi})}$, $R_{1i(\boldsymbol{\phi}\boldsymbol{\phi})}$ in b_i and $\widetilde{\mu}_{i(\boldsymbol{\phi})}$ in R_{2i} at $\boldsymbol{\phi} = \widehat{\boldsymbol{\phi}}$. However, their analytical expressions are too complicated to use them in practice. Thus we utilize the numerical derivatives which were suggested in Lahiri et al. (2007). Let $\{z_m\}$ be a sequence of positive real numbers converging to 0. Based on $\{z_m\}$, we define

$$\begin{aligned} f^*_{a(\phi_j)}(y_i, \widehat{\boldsymbol{\phi}}) &= \frac{1}{2z_m}\left\{f_a(y_i, \widehat{\boldsymbol{\phi}} + z_m \boldsymbol{e}_j) - f_a(y_i, \widehat{\boldsymbol{\phi}} - z_m \boldsymbol{e}_j)\right\}, \quad a = 1, 2 \\ \widetilde{\mu}^*_{i(\phi_j)}(y_i, \widehat{\boldsymbol{\phi}}) &= \frac{1}{2z_m}\left\{\widetilde{\mu}_i(y_i, \widehat{\boldsymbol{\phi}} + z_m \boldsymbol{e}_j) - \widetilde{\mu}_i(y_i, \widehat{\boldsymbol{\phi}} - z_m \boldsymbol{e}_j)\right\} \\ R^*_{1i(\phi_j)}(y_i, \widehat{\boldsymbol{\phi}}) &= \frac{1}{2z_m}\left\{R_{1i}(y_i, \widehat{\boldsymbol{\phi}} + z_m \boldsymbol{e}_j) - R_{1i}(y_i, \widehat{\boldsymbol{\phi}} - z_m \boldsymbol{e}_j)\right\} \end{aligned}$$

where $\boldsymbol{e}_j$ is a vector of 0's other than the j-th element is 1. Similarly, we define approximations of the second-order partial derivatives of R_{1i} as

$$\begin{aligned} R^*_{1i(\phi_j\phi_j)}(y_i, \widehat{\boldsymbol{\phi}}) =& \frac{1}{z_m^2}\left\{R_{1i}(y_i, \widehat{\boldsymbol{\phi}} + z_m \boldsymbol{e}_j) + R_{1i}(y_i, \widehat{\boldsymbol{\phi}} - z_m \boldsymbol{e}_j) - 2R_{1i}(y_i, \widehat{\boldsymbol{\phi}})\right\}, \quad j = 1, \ldots, k \\ R^*_{1i(\phi_j\phi_\ell)}(y_i, \widehat{\boldsymbol{\phi}}) =& \frac{1}{2z_m^2}\Big[\left\{R_{1i}(y_i, \widehat{\boldsymbol{\phi}} + z_m \boldsymbol{e}_{j\ell}) + R_{1i}(y_i, \widehat{\boldsymbol{\phi}} - z_m \boldsymbol{e}_{j\ell}) - 2R_{1i}(y_i, \widehat{\boldsymbol{\phi}})\right\} \\ & - z_m^2\left\{R^*_{1i(\phi_j\phi_j)}(y_i, \widehat{\boldsymbol{\phi}}) + R^*_{1i(\phi_\ell\phi_\ell)}(y_i, \widehat{\boldsymbol{\phi}})\right\}\Big], \quad j \neq \ell, \end{aligned}$$

where $\boldsymbol{e}_{j\ell} = \boldsymbol{e}_j + \boldsymbol{e}_\ell$. The justification of the approximations based on these numerical derivatives is given in the following theorem, where the proof is given in Section 8.6.

Theorem 8.3. *Under Assumption 8.1, we have*

$$|f^*_{a(\phi_j)}(y_i, \widehat{\boldsymbol{\phi}}) - f_{a(\phi_j)}(y_i, \widehat{\boldsymbol{\phi}})| = O_p(z_m), \qquad |\widetilde{\mu}^*_{i(\phi_j)}(y_i, \widehat{\boldsymbol{\phi}}) - \widetilde{\mu}_{i(\phi_j)}(y_i, \widehat{\boldsymbol{\phi}})| = O_p(z_m)$$
$$|R^*_{1i(\phi_j)}(y_i, \widehat{\boldsymbol{\phi}}) - R_{1i(\phi_j)}(y_i, \widehat{\boldsymbol{\phi}})| = O_p(z_m), \quad |R^*_{1i(\phi_j\phi_\ell)}(y_i, \widehat{\boldsymbol{\phi}}) - R_{1i(\phi_j\phi_\ell)}(y_i, \widehat{\boldsymbol{\phi}})| = O_p(z_m)$$

From Theorem 8.3, the second-order unbiasedness of the MSE estimator given in Theorem 8.2 is still valid as far as $z_m = o(m^{-1})$. In our numerical investigation given in the next section, we use $z_m = m^{-5/4}$.

8.4 Simulation Studies

8.4.1 Prediction error comparison

We first evaluated a finite sample performance of the proposed empirical uncertain Bayes method. Specifically, we compared the EUB estimator with the traditional empirical Bayes (EB) estimator. We focused on the two models: Poisson-gamma and binomial-beta models described in Section 8.2.3.

For the Poisson-gamma model, we considered the following data generating process:

$$\text{PG}: \quad (n_i y_i)|\theta_i \sim \text{Po}(n_i\theta_i), \quad \theta_i|(s_i = 1) \sim \text{Ga}(\nu \exp(\beta_0 + \beta_1 x_i), \nu), \quad P(s_i = 1) = p,$$

where $\beta_0 = 0, \beta_1 = 0.5$, $\nu = 5$, $m = 50$, n_i's were generated from the uniform distribution on $\{5, 6, \ldots, 30\}$, and x_i's were generated from a standard normal distribution. The prior probability p takes the values $0.2, 0.4, 0.6, 0.8$ and 1, where the conventional Poisson-gamma model corresponds to the data generating process with $p = 1$. We computed both the EUB and EB estimators from the simulated data set. Based on $R = 5,000$ iterations of the data generation, we calculated the mean squared error and the absolute bias which are respectively defined as

$$\text{MSE}_i = \frac{1}{R}\sum_{r=1}^{R}\left(\widehat{\theta}_i^{(r)} - \theta_i^{(r)}\right)^2, \quad \text{Bias}_i = \frac{1}{R}\Big|\sum_{r=1}^{R}\left(\widehat{\theta}_i^{(r)} - \theta_i^{(r)}\right)\Big|. \tag{8.18}$$

Define $\text{MSE}_i(\text{EUB})$ and $\text{MSE}_i(\text{EB})$ be the simulated values MSE_i for the EUB estimator and the EB estimator, respectively. Then we computed the ratio $\text{Ra}_i = \text{MSE}_i(\text{EUB})/\text{MSE}_i(\text{EB})$ for each i, and calculated the $q\%$ quantiles of $\{\text{Ra}_1, \ldots, \text{Ra}_m\}$ for $q = 5, 25, 50, 75$ and 95. Hence, if Ra_i is smaller than 1, the EUB estimator performs better than the EB estimator in terms of MSE. We similarly define the ratio of the absolute biases, and the results for the five p patterns are given in Table 8.1. In the scenario $p = 1$, the traditional Poisson-gamma model is the true model and uncertain model is overfitting. However, the results show that the EUB estimator performs as well as the EB estimator, which indicates that the effect of overfitting seems small. Moreover, from Table 8.1, when p is smaller than 1, it is revealed that the EUB estimator improve the EB estimator in terms of both the MSE and the absolute bias, and the improvement is greater as p gets smaller.

We next compared performances of the two estimators in the binomial-beta model using the data generating process:

$$\text{BB}: \quad (n_i y_i)|\theta_i \sim \text{Bin}(n_i, \theta_i), \quad \theta_i|(s_i = 1) \sim \text{Beta}(\nu m_i, \nu(1 - m_i)), \quad P(s_i = 1) = p,$$

with $m_i = \exp(\beta_0 + \beta_1 x_i)/\{1 + \exp(\beta_0 + \beta_1 x_i)\}$, where $\beta_0 = 0, \beta_1 = 0.5$, $\nu = 5$, $m = 50$, n_i's were generated from the uniform distribution on $\{10, 11, \ldots, 30\}$, and x_i's were generated from a standard normal distribution. Similarly to the previous study, we simulated the MSE and the absolute bias using (8.18) with $R = 5000$, and computed the quantiles of the ratios. The results are given in Table 8.1, which shows the similar results to the Poisson-gamma case. However, the amount of improvement seems smaller than that in the Poisson-gamma case, but the EUB estimator performs better than the EB estimator in the binomial-beta case.

Table 8.1: Simulated ratios of the MSEs and absolute biases of the EUB estimator over the EB estimator.

		MSE					Absolute bias				
	p	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
PG	0.2	0.156	0.355	0.542	0.717	0.842	0.554	0.628	0.707	0.832	0.878
	0.4	0.298	0.454	0.578	0.710	0.815	0.652	0.693	0.726	0.799	0.874
	0.6	0.540	0.640	0.728	0.826	0.894	0.795	0.821	0.843	0.876	0.911
	0.8	0.700	0.832	0.876	0.912	0.962	0.890	0.911	0.925	0.949	0.977
	1	0.985	0.993	1.000	1.005	1.011	0.989	0.993	0.999	1.002	1.007
BB	0.2	0.395	0.608	0.730	0.807	0.996	0.490	0.583	0.650	0.768	0.980
	0.4	0.488	0.736	0.827	0.887	0.983	0.721	0.747	0.805	0.862	0.987
	0.6	0.730	0.866	0.909	0.934	0.983	0.813	0.851	0.886	0.924	0.994
	0.8	0.865	0.946	0.970	0.984	0.993	0.916	0.938	0.961	0.971	0.992
	1	0.966	0.980	1.001	1.007	1.017	0.979	0.987	0.995	1.004	1.012

8.4.2 Sensitivity to distributional assumptions

We next investigated sensitivity to distributional assumptions in the proposed model. Here we focused on the Poisson-gamma model as considered in the previous simulation study:

$$(n_i y_i)|\theta_i \sim \mathrm{Po}(n_i\theta_i), \quad \theta_i|(s_i = 1) \sim \mathrm{Ga}(\nu \exp(\beta_0 + \beta_1 x_i), \nu), \quad P(s_i = 1) = p,$$

where $p = 0.5$, and other settings β, β_1, ν, n_i and x_i are set as the same values as in the previous section. To asses sensitivity of the distributional assumption of the proposed method, we consider the two alternative distributions: a log-normal distribution and a two-point distribution for θ_i instead of the gamma distribution. Noting that $\mathrm{E}[\theta_i] = m_i$ and $\mathrm{Var}(\theta_i) = m_i/\nu$ under the gamma distribution, we scaled two distributions to have the same expectation and variance. Specifically, we set $\log\theta_i \sim N(\log(m_i/\sqrt{1 + 1/\nu m_i}), \log(1 + 1/\nu m_i))$ for the log-normal distribution, and $P(\theta_i = m_i + \sqrt{m_i/\nu}) = P(\theta_i = m_i - \sqrt{m_i/\nu}) = 0.5$ for the two-point distribution. Based on $R = 5000$ simulation runs, we computed the MSE and absolute bias with the formula (8.18) for three underlying distributions. In Table 8.2, we show the five quantiles of the simulated MSE and absolute bias. It is observed that both the MSE and the absolute bias in the misspecified cases of log-normal and two-point distributions are larger than the correctly specified case of the gamma distribution. The absolute biases in the two

misspecified cases are about twice as large as that in the gamma case, so that the inflation of the absolute bias seems relatively large. However, the difference in the MSE is around 10%, so that the effect of misspecification on MSE seems relatively small.

Table 8.2: Quantiles of the simulated MSE and absolute bias of the EUB estimator under the three underlying distributions (the values are multiplied by 100).

	MSE					Absolute bias				
distribution	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
Gamma	1.31	2.55	3.98	5.92	11.06	0.03	0.10	0.25	0.48	0.71
Log-normal	1.32	2.47	4.15	5.85	10.89	0.05	0.33	0.49	0.72	1.25
two-point	1.57	2.87	4.50	6.46	10.58	0.08	0.28	0.51	0.84	1.18

8.4.3 Finite sample performance of the CMSE estimator

Finally we investigated a finite sample behavior of the CMSE estimator provided in Theorem 8.2 in the UPG model. We considered the simple data generating process without covariates:

$$(n_i y_i)|\lambda_i \sim \mathrm{Po}(n_i\lambda_i), \quad \lambda_i|(s_i = 1) \sim \mathrm{Ga}(\nu\exp(\beta), \nu), \quad P(s_i = 1) = p, \tag{8.19}$$

with $\beta = 1$, $\nu = 5$, $p = 0.5$ and $n_i = 10$. For the number of areas, we consider the two cases of $m = 50$ and $m = 100$. For conditioning values of y_α, we consider α-quantiles of the marginal distribution of y_i, where $\alpha = 0.1, 0.2, \ldots, 0.9$, and calculate these values by generating $10,000$ random samples from (8.19). To get the simulated values of the CMSE given y_α, we generate random samples from (8.19) and replace y_1 with y_α, and we computed the EUB estimator of $\widehat{\lambda}_1$. For the true values of λ_1, since y_α is given, we generate λ_1 from the posterior distribution $\lambda_1|y_\alpha \sim r_1\mathrm{Ga}(\widetilde{\lambda}_1, (n_1+\nu)^{-1}\widetilde{\lambda}_1) + (1-r_1)\delta_{\lambda_1}(\exp(\beta))$ with $\widetilde{\lambda}_1 = (n_1+\nu)^{-1}(n_1 y_\alpha + \nu\exp(\beta))$ and r_1 given in (8.12). Then, based on $R = 10,000$ iteration, we calculate the simulated values of the CMSE defined as

$$\mathrm{CM}_\alpha = \frac{1}{R}\sum_{r=1}^{R}(\widehat{\lambda}_1^{(r)} - \lambda_1^{(r)})^2,$$

where $\widehat{\lambda}_1^{(r)}$ and $\lambda_1^{(r)}$ are the EUB estimates of λ_1, and $\lambda_1^{(r)}$ is the generated value from the distribution of $\lambda_1|y_\alpha$ in the rth iteration.

For evaluation of the CMSE estimator, we generated random samples from (8.19) and replace y_1 with y_α, and get CMSE estimators with $B = 100$ bootstrap samples and $z_m = m^{-5/4}$ for computing the numerical derivatives. This procedure is repeated $S = 2,000$ times and calculated the percentage relative bias (RB) and the coefficient of variation (CV) defined as

$$\mathrm{RB}_\alpha = \frac{1}{S}\sum_{s=1}^{S}\left(\frac{\widehat{\mathrm{CM}}_\alpha - \mathrm{CM}_\alpha}{\mathrm{CM}_\alpha}\right)\times 100, \quad \mathrm{CV}_\alpha = \sqrt{\frac{1}{S}\sum_{s=1}^{S}\left(\frac{\widehat{\mathrm{CM}}_\alpha - \mathrm{CM}_\alpha}{\mathrm{CM}_\alpha}\right)^2}.$$

Remember that the suggested CMSE estimator given in Theorem 8.2 is second-order unbiased. To emphasize the importance of bias correction in estimating the CMSE, we also computed the two criteria of the naive CMSE estimator defined as $\widehat{\mathrm{CM}}_{\alpha(N)} = R_{11}(y_\alpha, \widehat{\boldsymbol{\phi}})$. It is noted that the naive estimator has the first order bias, because it ignores the second term R_{2i} and the bias of the plug-in estimator of R_{1i}. For the naive estimator, we calculated RB and CV based on the same number of iteration and we define RBN and CVN as RB and CV of the naive estimator. The resulting values are given in Table 8.3 for both $m = 50$ and $m = 100$.

It is observed from Table 8.3, the naive estimator has the serious negative bias when $m = 50$. Especially, when the condition values are upper or lower quantiles, the negative bias tends to be larger. This comes from the fact that the naive estimator ignores the positive $O_p(m^{-1})$ term in the CMSE decomposition given in Theorem 8.1. Since practitioners decide policies or investments based on estimated values as well as their risk estimates, the underestimation of the CMSE is considered serious in practice. Hence, the results in Table 8.3 show that the naive CMSE estimator without bias correction is not suitable for practical use. On the other hand, the bias-corrected CMSE estimator works well in both $m = 50$ and $m = 100$ and provides accurate estimation of the CMSE in terms of the relative biases. Concerning the CV values, the bias-corrected estimator has a slightly larger CV than the naive estimator in most cases. This is because the bias corrected terms increase the variance of the estimator. However, the difference is not so significant. Thus, the bias-corrected CMSE estimator is useful in practice.

Table 8.3: Percentage of relative bias and coefficient of variation.

	α	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
	y_i	1.8	2.1	2.3	2.5	2.7	2.9	3.0	3.3	3.6
$m = 50$	RB	1.21	-1.01	6.87	6.56	5.87	3.43	2.42	-3.95	-9.59
	RBN	-27.7	-23.5	-13.8	-6.56	-3.28	-3.86	-8.37	-21.1	-28.6
	CV	0.53	0.36	0.53	0.67	0.73	0.68	0.64	0.43	0.34
	CVN	0.38	0.37	0.43	0.53	0.61	0.55	0.51	0.39	0.39
$m = 100$	RB	-0.23	-0.45	3.28	3.68	2.02	-0.12	0.55	-2.94	-5.82
	RBN	-16.9	-12.9	-6.33	-3.75	4.79	0.97	-2.55	-12.4	-17.9
	CV	0.28	0.25	0.41	0.57	0.63	0.58	0.52	0.32	0.24
	CVN	0.28	0.27	0.35	0.49	0.52	0.49	0.42	0.29	0.27

8.5 Illustrative Examples

8.5.1 Historical mortality data in Tokyo

The mortality rate is a representative index in demographics and has been used in various fields. Especially, in economic history, one can discover new knowledge from a spatial distribution of mortality rate in small areas. As divisions get smaller (e.g. city→town→block...), one can get a more informative spatial distribution. However, the direct estimate of the mortality rate in small area with extremely low population has high variability, which may leads

to incorrect recognition of the spatial distribution. Therefore, it is desirable to use smoothed and stabilized estimates through empirical Bayes methods.

We here focus on the mortality data in Tokyo, 1930. The data set consists of the observed mortalities z_i and the number of population N_i in the ith area in Tokyo. Such area-level data are available for $m = 1,371$ small areas. We first computed the expected mortality in the ith area as $n_i = N_i \sum_{j=1}^m z_j / \sum_{j=1}^m N_j$. The standardized mortality ratio (SMR) is defined as the ratio of the actual mortality to the expected mortality for each area, which is often used in epidemiology as an indicator of potential mortality risk. Then, the direct estimator of the SMR in the ith area is $y_i = z_i/n_i$. It is noted that $y_i = 0$ in 84 areas, the number of areas with SMR larger than 1 is 526, and the maximum value of y_i is 16.4.

For this data set, we apply the two models: the uncertain Poisson-gamma model described Section 8.2.3 and the traditional Poisson-gamma model, described as

$$\begin{aligned} \text{UPG}: \ & n_i y_i | \lambda_i \sim \text{Po}(n_i \lambda_i), \quad \lambda_i | s_i \sim \text{Ga}(\nu \exp(\beta), \nu), \quad P(s_i = 1) = p \\ \text{PG}: \ & n_i y_i | \lambda_i \sim \text{Po}(n_i \lambda_i), \quad \lambda_i \sim \text{Ga}(\nu \exp(\beta), \nu), \end{aligned}$$

where $\lambda_i = E(y_i|\lambda_i)$ denotes the 'true' SMR in the i-th area, which we want to estimate. Using the EM algorithm in Section 8.2.2 with 5000 Monte Calro samples in each E-step, we get the point estimates of the parameters of the two models as shown in Table 8.4. For comparison of the two models, we computed AIC and BIC based on the maximum marginal likelihood, and the results are also given in Table 8.4. In terms of AIC and BIC, the proposed UPG model fits better than the traditional PG model for this data set. This comes from the feature of the data. In the upper left panel of Figure 8.2, we show the sample plot of the expected mortality n_i and the SMR y_i, noting that the solid line corresponds to the estimated regression line $y_i = \exp(\widehat{\beta})$ in the UPG model. It is observed that most y_i are distributed around the regression line, and the random area effects are necessary in most areas. The UPG model tells us about the feature of the data through the estimate of p. The lower left panel of Figure 8.2 provides a scatter plot of the estimated conditional probability $P(s_i = 1|y_i)$ and the SMR y_i, where the conditional probability $P(s_i = 1|y_i)$ corresponds to the probability of existing random area effect in the ith area when y_i is observed. The solid line corresponds to the estimated regression line $y_i = \exp(\widehat{\beta})$ in the UPG model. From the figure, we can see that the estimates of $P(s_i = 1|y_i)$ are dramatically different from area to area, and the probability gets lower as SMR is closer to the regression line. To see the difference of estimated values of λ_i, in the upper right panel of Figure 8.2, we present the relative differences between estimators from the two models, which are defined as $(\widehat{\lambda}_i^{\text{UPG}} - \widehat{\lambda}_i^{\text{PG}})/\widehat{\lambda}_i^{\text{PG}}$, where $\widehat{\lambda}_i^{\text{UPG}}$ and $\widehat{\lambda}_i^{\text{PG}}$ are empirical Bayes estimates of λ_i from the UPG model and the PG model, respectively. We can observe that the differences are around 10% and are not negligible.

We next calculated the CMSE estimates of the EUB estimates $\widehat{\lambda}_i^{\text{UPG}}$ using Theorem 8.2 with $B = 100$ and $z_m = m^{-5/4}$. For comparison, we also computed the CMSE estimates of $\widehat{\lambda}_i^{\text{PG}}$ using Theorem 8.2 with $p = 1$, $B = 100$ and $z_m = m^{-5/4}$. Then, we computed their difference and their histogram over areas is given in the lower right panel of Figure 8.2. In the figure, the positive value indicates that the EUB estimator has the smaller CMSE value than the EB estimator, and it is revealed that the EUB estimator can improve the estimation risk over the EB estimator in many areas. In particular, the mean values of CMSEs are 4.2×10^{-2} for UPG and 5.4×10^{-2} for PG, so that the EUB estimator can improve 20% CMSE values over the traditional EB estimator on average.

Finally, we assessed the performance of the two models in terms of prediction accuracy in non-sampled areas. Since areas with small n_i have high variability, we consider to predict y_i of areas with n_i larger than the α-quantile of n_i's, denoted by q_α. Thus we omitted areas with n_i larger than q_α and computed the estimates of the model parameters using the remaining data. For fixed α, we define the predictive criterion (PC) as

$$\mathrm{PC}_\alpha = \sum_{i=1}^{m} I(n_i > q_\alpha)\,(\widehat{m}_i - y_i)^2 \Big/ \sum_{i=1}^{m} I(n_i > q_\alpha), \tag{8.20}$$

noting that $\widehat{m}_i$ is the best predictor in non-sampled areas. In this example, $\widehat{m}_i = \exp(\widehat{\beta})$. The values of PC were computed for three quantiles of $\alpha = 0.90, 0.95$ and 0.99 and reported in Table 8.4. It is revealed that the EUB method can improve PC values over the EB method by about 10%.

Table 8.4: Point estimates of the model parameters and values of AIC, BIC and PC (multiplied by 100) for the three thresholds.

Estimates	$\widehat{\beta}$	$\widehat{\nu}$	$\widehat{p}$	AIC	BIC	$\mathrm{PC}_{0.90}$	$\mathrm{PC}_{0.95}$	$\mathrm{PC}_{0.99}$
UPG	-0.039	5.15	0.56	8142.17	8157.84	8.00	7.70	5.69
PG	-0.052	7.42	—	8265.12	8275.57	8.67	8.24	6.15

8.5.2 *Poverty rates in Spanish provinces*

We next applied our method to the income data set in Spanish provinces as used in Section 3.4. In this application, we focus on estimating area-level poverty rates. We set the poverty level as 0.7 times the median of all the observed incomes, and computed the direct estimates of the poverty rates. As covariates, we calculated area-level rates of female and labors. The scatter plot of the pairs (n_i, y_i) is given in the left panel of Figure 8.3, from which we can observe that the direct estimate y_i has higher variability as n_i gets smaller.

For the data set, we applied the two models: the uncertain binomial-beta (UBB) model and the traditional binomial-beta (BB) model, described as

$$\begin{aligned} \text{UBB}: &\ n_i y_i|\theta_i \sim \mathrm{Bin}(n_i, \theta_i), \quad \theta_i|s_i \sim \mathrm{Beta}(\nu m_i, \nu(1-m_i)), \quad P(s_i = 1) = p \\ \text{BB}: &\ n_i y_i|\theta_i \sim \mathrm{Bin}(n_i, \theta_i), \quad \theta_i \sim \mathrm{Beta}(\nu m_i, \nu(1-m_i)), \end{aligned}$$

where y_i is the direct estimate of the true poverty rate p_i, n_i is the number of observations in the ith area, $m_i = \mathrm{logit}(\beta_0 + \beta_1 g_i + \beta_2 f_i)$ for $\mathrm{logit}(x) = \exp(x)/(1+\exp(x))$, and g_i and f_i are rates of populations of females and labors, respectively. The point estimates of the model parameters based on the EM algorithm in Section 8.2.2 with 5000 Monte Carlo samples are shown in Table 8.5. The signs of $\widehat{\beta}_1$ and $\widehat{\beta}_2$ are reasonable. From the table, it is observed that the estimate of p in the UBB model is almost 1, which implies that the traditional BB model is appropriate for this data set. Actually, the values of AIC and BIC based on the marginal likelihood, given in Table 8.5, support the BB model rather than the UBB model.

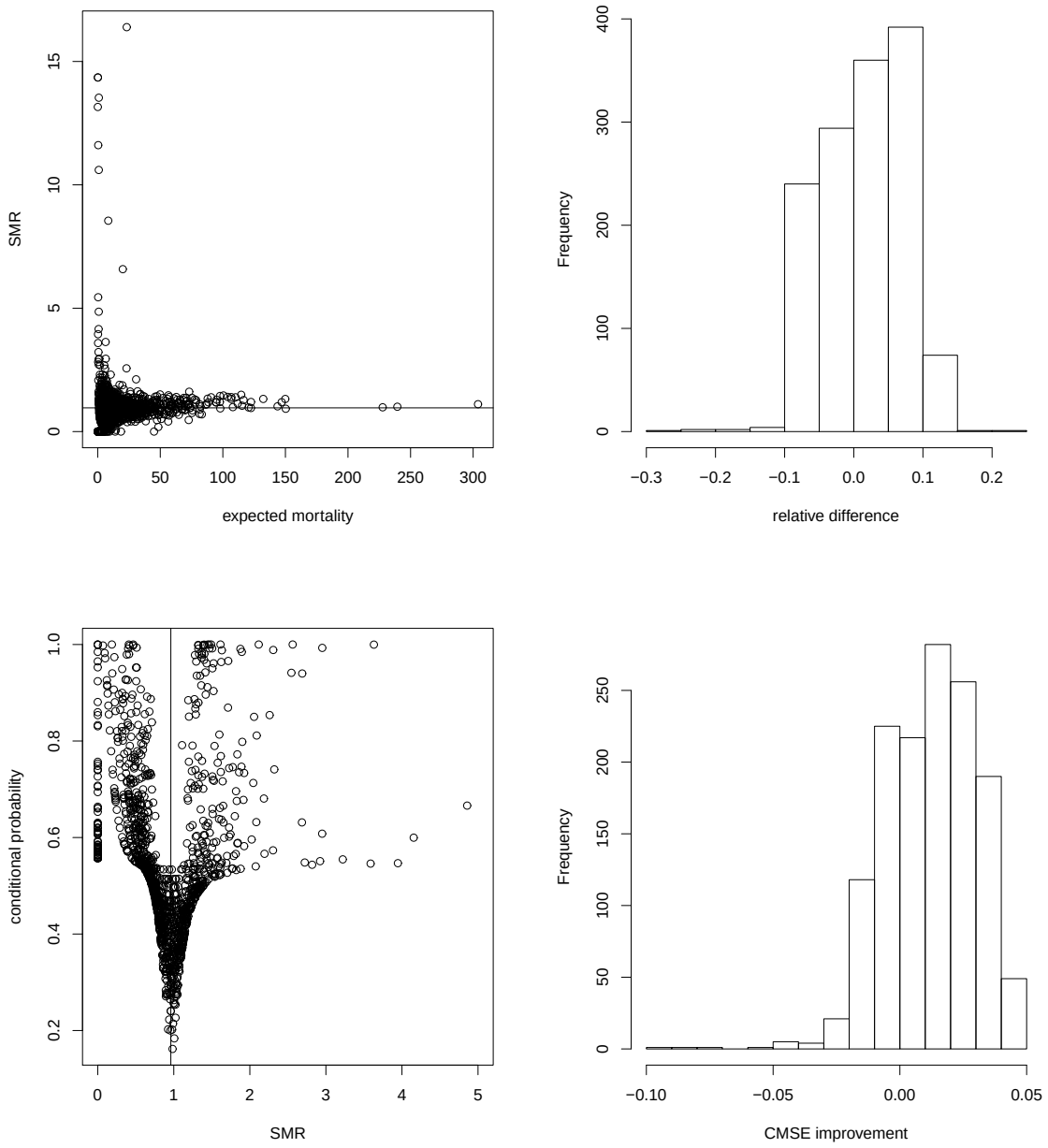

Figure 8.2: Sample plot of the expected mortality n_i and the SMR y_i (upper-left), the scaled difference of two predictors: $(\widehat{\lambda}_i^{\mathrm{UPG}} - \widehat{\lambda}_i^{\mathrm{PG}})/\widehat{\lambda}_i^{\mathrm{PG}}$ (upper-right), sample plot of the estimates of conditional probability $P(s_i = 1|y_i)$ and the SMR y_i (lower-left), and histogram of improvement of estimated CMSE: $\widehat{\mathrm{CM}}_i^{\mathrm{PG}} - \widehat{\mathrm{CM}}_i^{\mathrm{UPG}}$ (lower-right).

Concerning the differences of predicted values, we provide in the right panel of Figure 8.3 the histogram of the relative differences: $(\widehat{\theta}_i^{\rm UBB} - \widehat{\theta}_i^{\rm BB})/\widehat{\theta}_i^{\rm BB}$, where $\widehat{\theta}_i^{\rm UBB}$ and $\widehat{\theta}_i^{\rm BB}$ are predicted values from the UBB and the BB models. It shows that the differences are smaller than 1% in most areas.

We next calculated the CMSE estimates of the EUB and the EB estimates using Theorem 8.2 with $B = 100$ and $z_m = m^{-5/4}$. These two estimates are expected to be similar, but the CMSE estimates of the EUB estimates are negative in some areas, while those of the EB estimates are all positive. This may comes from the instability of estimating p close to 1.

Finally, we considered the performances of prediction for non-sampled areas. Similarly to the previous section, we considered the predictive criterion (PC) defined in (8.20) with $\widehat{m}_i = \text{logit}(\widehat{\beta}_0 + \widehat{\beta}_1 g_i + \widehat{\beta}_2 f_i)$. The values of PC were computed for $\alpha = 0.70, 0.80$ and 0.90 and reported in Table 8.5. It is observed that the UBB model provides the performance better than the BB model while the differences are quite small.

Table 8.5: Point estimates of the model parameters, values of AIC, BIC and PC (multiplied by $1,000$) for the three thresholds.

Estimates	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\nu}$	$\widehat{p}$	AIC	BIC	$\text{PC}_{0.70}$	$\text{PC}_{0.80}$	$\text{PC}_{0.90}$
UBB	-1.92	2.91	-1.03	41.33	0.96	459.67	469.42	5.59	5.42	5.17
BB	-2.14	3.36	-1.07	42.93	—	457.74	465.55	5.61	5.42	5.19

8.6 Technical Issues

8.6.1 Checking Assumption 8.1 in typical three models.

(Fay-Herriot model). It follows from (8.10) that

$$f_{1(\boldsymbol{\beta})}(y_i;\boldsymbol{\phi}) = f_1(y_i;\boldsymbol{\phi})\left(\frac{y_i - \boldsymbol{x}_i^t\boldsymbol{\beta}}{A + D_i}\right)\boldsymbol{x}_i, \quad f_{1(A)}(y_i;\boldsymbol{\phi}) = \frac{f_1(y_i;\boldsymbol{\phi})}{2(A+D_i)^2}\left\{\left(y_i - \boldsymbol{x}_i^t\boldsymbol{\beta}\right)^2 - A - D_i\right\}.$$

Using $f_1(y_i;\boldsymbol{\phi}) \le 1/\sqrt{2\pi A}$, we can see that $|f_{1(\phi_j)}(y_i;\boldsymbol{\phi})|$, $|f_{1(\phi_j\phi_\ell)}(y_i;\boldsymbol{\phi})|$ and $|f_{1(\phi_j\phi_\ell\phi_k)}(y_i;\boldsymbol{\phi})|$ can be evaluated from above by 6th order polynomials of y_i and the assumption (iii) is satisfied for $a = 1$. The case of $a = 2$ can be shown similarly.

(Poisson-gamma model). It is noted that $f_{1(\phi_k)}(y_i;\boldsymbol{\phi}) = f_1(y_i;\boldsymbol{\phi})\partial \log f_1(y_i;\boldsymbol{\phi})/\partial\phi_k$. From (8.11), it holds that

$$\frac{\partial \log f_1(y_i;\boldsymbol{\phi})}{\partial \beta_k} = \nu x_{ik} m_i \left\{\psi(n_i y_i + \nu m_i) - \psi(\nu m_i)\right\}, \quad k = 1, \ldots, q,$$
$$\frac{\partial \log f_1(y_i;\boldsymbol{\phi})}{\partial \nu} = m_i \left\{\psi(n_i y_i + \nu m_i) - \psi(\nu m_i)\right\} - \frac{n_i y_i}{n_i + \nu} + m_i\left\{\frac{n_i}{n_i+\nu} + \log\left(\frac{n_i}{n_i+\nu}\right)\right\},$$

where $\psi(\cdot)$ is the digamma function $\psi(x) = \mathrm{d}\log\Gamma(x)/\mathrm{d}x$. Using the fact that $\psi(x) \approx \log x$ for large x, we have $|\partial \log f_1(y_i;\boldsymbol{\phi})/\partial\beta_k| = O_p(\log y_i)$ and $|\partial \log f_1(y_i;\boldsymbol{\phi})/\partial\nu| = O_p(y_i)$ for

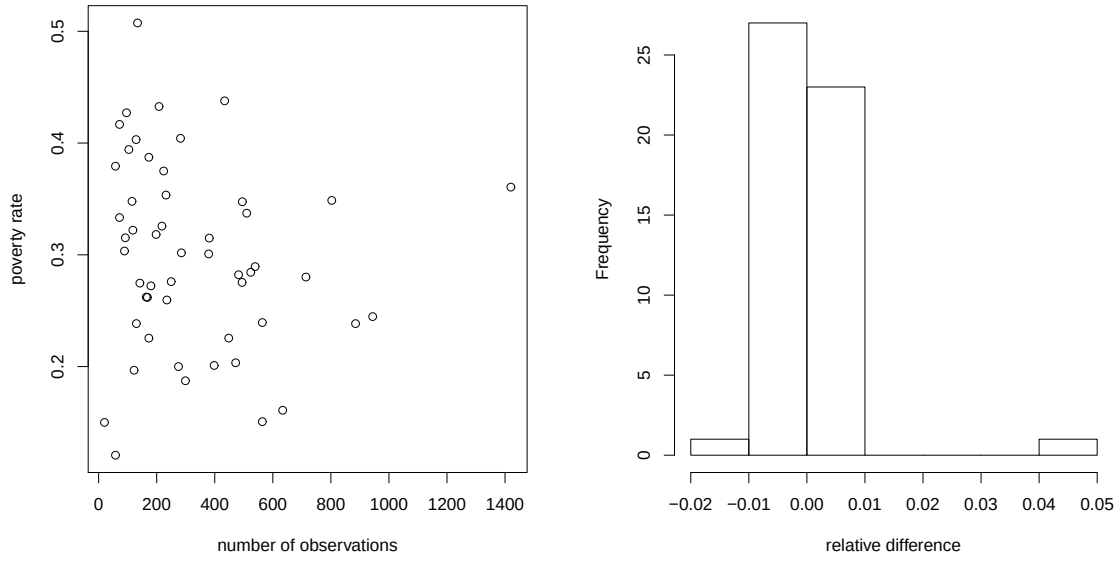

Figure 8.3: The scatter plot of the number of observations n_i and the direct estimate of poverty rate y_i (left), and the histogram of relative difference $(\widehat{\theta}_i^{\mathrm{UBB}} - \widehat{\theta}_i^{\mathrm{BB}})/\widehat{\theta}_i^{\mathrm{BB}}$ (right).

large y_i. Since there exists $c \in \mathbb{R}$ such that $f_1(y_i;\boldsymbol{\phi}) \leq c$, $|f_{1(\phi_k)}(y_i;\boldsymbol{\phi})|$ is bounded above by an liner function of y_i. Concerning the second derivatives, we note that

$$f_{1(\phi_k\phi_\ell)}(y_i;\boldsymbol{\phi}) = f_1(y_i;\boldsymbol{\phi}) \left\{ \frac{\partial^2 \log f_1(y_i;\boldsymbol{\phi})}{\partial\phi_k\partial\phi_\ell} + \frac{\partial \log f_1(y_i;\boldsymbol{\phi})}{\partial\phi_k} \frac{\partial \log f_1(y_i;\boldsymbol{\phi})}{\partial\phi_\ell} \right\}. \tag{8.21}$$

Moreover, the straightforward calculation shows that

$$\frac{\partial^2 \log f_1(y_i;\boldsymbol{\phi})}{\partial\beta_k^2} = \nu x_{ik}^2 m_i \left\{\psi(n_i y_i + \nu m_i) - \psi(\nu m_i)\right\} + \nu x_{ik}^2 m_i^2 \left\{\psi^{(1)}(n_i y_i + \nu m_i) - \psi^{(1)}(\nu m_i)\right\},$$

where $\psi^{(n)}(x) = \mathrm{d}^n\psi(x)/\mathrm{d}x^n$ is a polygamma function. Since $\psi^{(n)}(x) \approx (-1)^n(n-1)!x^{-n}$ for large x, we have $|\partial^2 \log f_1(y_i;\boldsymbol{\phi})/\partial\beta_k^2| = O(\log y_i)$ for large y_i. Similarly, we obtain $|\partial^2 \log f_1(y_i;\boldsymbol{\phi})/\partial\beta_k\partial\nu| = O(\log y_i)$ and $|\partial^2 \log f_1(y_i;\boldsymbol{\phi})/\partial\nu^2| = O(y_i)$ for large y_i. Thus from expression (8.21), $|f_{1(\phi_k\phi_\ell)}(y_i;\boldsymbol{\phi})|$ is bounded above by an quadratic function of y_i. The similar argument shows that $|f_{1(\phi_k\phi_\ell)}(y_i;\boldsymbol{\phi})|$ is bounded above by an cubic function of y_i. Hence, conditional (iii) is satisfied for $a = 1$, because $\mathrm{E}[y_i^c] < \infty$ for all $c > 0$ when y_i has the Poisson-gamma model. The case of $a = 2$ can be shown similarly.

(Binomial-beta model). Note that $f_1(y_i;\boldsymbol{\phi})$ and $f_2(y_i;\boldsymbol{\phi})$ have compact supports and the derivatives $f_{a(\phi_j)}(y_i;\boldsymbol{\phi})$, $f_{a(\phi_j\phi_\ell)}(y_i;\boldsymbol{\phi})$ and $f_{a(\phi_j\phi_\ell\phi_k)}(y_i;\boldsymbol{\phi})$ are finite for an interior point $\boldsymbol{\phi}$. Then condition (iii) is easy to check.

8.6.2 Proof of Theorem 8.3.

Let us fix $\boldsymbol{\phi}_0$ as an interior point of Φ. We here use the notation $C(y_i)$ as a generic function of y_i with $C(y_i) = O_p(1)$, and the notations bu_{1i} and $u_{2i} \in [-1, 1]$ as generic constants. Expanding $f_a(y_i, \boldsymbol{\phi}_0 + z_m \boldsymbol{e}_j)$ and $f_a(y_i, \boldsymbol{\phi}_0 - z_m \boldsymbol{e}_j)$ around $\boldsymbol{\phi}_0$, we get

$$f_a(y_i, \boldsymbol{\phi}_0 + z_m \boldsymbol{e}_j) = f_a(y_i, \boldsymbol{\phi}_0) + f_{a(\phi_j)}(y_i, \boldsymbol{\phi}_0) z_m + \frac{1}{2} f_{a(\phi_j \phi_\ell)}(y_i, \boldsymbol{\phi}_0 + u_{1i} z_m \boldsymbol{e}_j) z_m^2$$
$$f_a(y_i, \boldsymbol{\phi}_0 - z_m \boldsymbol{e}_j) = f_a(y_i, \boldsymbol{\phi}_0) - f_{a(\phi_j)}(y_i, \boldsymbol{\phi}_0) z_m + \frac{1}{2} f_{a(\phi_j \phi_\ell)}(y_i, \boldsymbol{\phi}_0 + u_{1i} z_m \boldsymbol{e}_j) z_m^2,$$

so that it follows that

$$\begin{aligned}
&(2z_m)|f^*_{a(\phi_j)}(y_i, \boldsymbol{\phi}_0) - f_{a(\phi_j)}(y_i, \boldsymbol{\phi}_0)| \\
&= |f_a(y_i, \boldsymbol{\phi}_0 + z_m \boldsymbol{e}_j) - f_a(y_i, \boldsymbol{\phi}_0 - z_m \boldsymbol{e}_j) - 2z_m f_{a(\phi_j)}(y_i, \boldsymbol{\phi}_0)| \\
&= \frac{1}{2} z_m^2 |f_{a(\phi_j \phi_\ell)}(y_i, \boldsymbol{\phi}_0 + u_{1i} z_m \boldsymbol{e}_j) - f_{a(\phi_j \phi_\ell)}(y_i, \boldsymbol{\phi}_0 + u_{2i} z_m \boldsymbol{e}_j)| \leq C(y_i) z_m^2,
\end{aligned}$$

from (iii) of Assumption 8.1. This shows the first part of Theorem 8.3.

To show the other parts, we prove that there exist functions $C_a(y_i) = O_p(1)$ for $a = 1, 2, 3$ such that

$$|r_{i(\phi_j)}(y_i, \boldsymbol{\phi}_0)| \leq C_1(y_i), \quad |r_{i(\phi_j \phi_\ell)}(y_i, \boldsymbol{\phi}_0)| \leq C_2(y_i), \quad |r_{i(\phi_j \phi_\ell \phi_k)}(y_i, \boldsymbol{\phi}_0)| \leq C_3(y_i). \tag{8.22}$$

The straightforward calculation shows that

$$\begin{aligned}
&r_{i(p)} = f_1 f_2 \{p f_1 + (1-p) f_2\}^{-2}, \quad r_{i(pp)} = -2 f_1 f_2 \{p f_1 + (1-p) f_2\}^{-3} (f_1 - f_2), \\
&r_{i(ppp)} = 6 f_1 f_2 \{p f_1 + (1-p) f_2\}^{-4} (f_1 - f_2)^2,
\end{aligned}$$

which are all bounded above by $C(y_i)$ since $f_a / (p f_1 + (1-p) f_2) \leq \max(p^{-1}, (1-p)^{-1})$ for $a = 1, 2$. Moreover, it is noted that

$$|r_{i(\phi_j)}| = \frac{p(1-p) \left| f_{1(\psi_j)} f_2 - f_1 f_{2(\psi_j)} \right|}{\{p f_1 + (1-p) f_2\}^2} \leq \frac{p |f_{1(\psi_j)}| + (1-p) |f_{2(\psi_j)}|}{p f_1 + (1-p) f_2} \leq C(y_i)$$

under (iii) of Assumption 8.1. Similarly, it can be shown that the higher order derivatives $r_{i(\phi_j \phi_\ell)}$, $r_{i(\phi_j \phi_\ell \phi_k)}$, $r_{i(p \phi_\ell)}$, $r_{i(pp \phi_\ell)}$ and $r_{i(p \phi_\ell \phi_k)}$ have the form $h(y_i, \boldsymbol{\phi}_0) / \{p f_1 + (1-p) f_2\}^c$, where c is a positive integer and $h(y_i, \boldsymbol{\phi}_0)$ is a polynomial of f_a, $f_{a(\phi_j)}$, $f_{a(\boldsymbol{\phi}_j \phi_k)}$ and $f_{a(\phi_j \phi_\ell \phi_k)}$, so that there exists $h^\dagger(y_i) = O_p(1)$ such that $h(y_i, \boldsymbol{\phi}_0) \leq h^\dagger(y_i)$. This establishes property (8.22). Using the property, we have

$$\begin{aligned}
&(2z_m) \left| \widetilde{\mu}^*_{i(\phi_j)}(y_i, \boldsymbol{\phi}_0) - \widetilde{\mu}_{i(\phi_j)}(y_i, \boldsymbol{\phi}_0) \right| \\
&\qquad = \frac{1}{2} z_m^2 \left| \widetilde{\mu}_{i(\phi_j \phi_j)}(y_i, \boldsymbol{\phi}_0 + u_{1i} z_m \boldsymbol{e}_j) - \widetilde{\mu}_{i(\phi_j \phi_j)}(y_i, \boldsymbol{\phi}_0 + u_{2i} z_m \boldsymbol{e}_j) \right| \leq C(y_i) z_m^2
\end{aligned}$$

and

$$\begin{aligned}
&(2z_m) \left| R^*_{1i(\phi_j)}(y_i, \boldsymbol{\phi}_0) - R_{1i(\phi_j)}(y_i, \boldsymbol{\phi}_0) \right| \\
&\qquad = \frac{1}{2} z_m^2 \left| R_{1i(\phi_j \phi_j)}(y_i, \boldsymbol{\phi}_0 + u_{1i} z_m \boldsymbol{e}_j) - R_{1i(\phi_j \phi_j)}(y_i, \boldsymbol{\phi}_0 + u_{2i} z_m \boldsymbol{e}_j) \right| \leq C(y_i) z_m^2.
\end{aligned}$$

Finally, we consider the approximation of the second-order partial derivatives of R_{1i}. Expanding $R_{1i}(\boldsymbol{\phi}_0 + z_m \boldsymbol{e}_j)$ and $R_{1i}(\boldsymbol{\phi}_0 - z_m \boldsymbol{e}_j)$ up to $O(z_m^3)$, we have

$$\begin{aligned} & z_m^2 \big| R^*_{1i(\phi_j\phi_j)}(y_i, \boldsymbol{\phi}_0) - R_{1i(\phi_j\phi_j)}(y_i, \boldsymbol{\phi}_0) \big| \\ & \quad = \frac{1}{6} z_m^3 \big| R_{1i(\phi_j\phi_j\phi_j)}(y_i, \boldsymbol{\phi}_0 + u_{1i} z_m \boldsymbol{e}_j) - R_{1i(\phi_j\phi_j\phi_j)}(y_i, \boldsymbol{\phi}_0 + u_{2i} z_m \boldsymbol{e}_j) \big| \leq C(y_i) z_m^3, \end{aligned}$$

from property (8.22). From this result, we obtain for $j \neq \ell$,

$$\begin{aligned} R^*_{1i(\phi_j\phi_\ell)}(y_i, \boldsymbol{\phi}_0) = \frac{1}{2z_m^2} \Big[& \{ R_{1i}(y_i, \boldsymbol{\phi}_0 + z_m \boldsymbol{e}_{j\ell}) + R_{1i}(y_i, \boldsymbol{\phi}_0 - z_m \boldsymbol{e}_{j\ell}) - 2R_{1i}(y_i, \boldsymbol{\phi}_0) \} \\ & - z_m^2 \Big\{ R_{1i(\phi_j\phi_j)}(y_i, \boldsymbol{\phi}_0) + R_{1i(\phi_\ell\phi_\ell)}(y_i, \boldsymbol{\phi}_0) \Big\} \Big] + O_p(z_m). \end{aligned}$$

It is noted that

$$\begin{aligned} R_{1i}(y_i, & \boldsymbol{\phi}_0 + z_m \boldsymbol{e}_{j\ell}) = R_{1i}(y_i, \boldsymbol{\phi}_0) + R_{1i(\psi_j)}(y_i, \boldsymbol{\phi}_0) z_m + R_{1i(\phi_\ell)}(y_i, \boldsymbol{\phi}_0) z_m + R_{1i(\phi_j\phi_\ell)}(y_i, \boldsymbol{\phi}_0) z_m^2 \\ & + \frac{1}{2} R_{1i(\phi_j\phi_j)}(y_i, \boldsymbol{\phi}_0) z_m^2 + \frac{1}{2} R_{1i(\phi_\ell\phi_\ell)}(y_i, \boldsymbol{\phi}_0) z_m^2 + \frac{1}{6} z_m^3 \sum_{j,\ell,k} R_{1i(\phi_j\phi_\ell\phi_k)}(y_i, \boldsymbol{\phi}_0 + z_m u_{1j\ell k} \boldsymbol{e}_{j\ell k}) \end{aligned}$$

where $u_{1j\ell k} \in [-1, 1]$ and $\boldsymbol{e}_{j\ell k} = \boldsymbol{e}_j + \boldsymbol{e}_\ell + \boldsymbol{e}_k$. Then it follows that

$$\begin{aligned} & z_m^2 | R^*_{1i(\phi_j\phi_\ell)}(y_i, \boldsymbol{\phi}_0) - R_{1i(\phi_j\phi_\ell)}(y_i, \boldsymbol{\phi}_0) \big| \\ & \quad = \frac{1}{6} z_m^3 \Big| \sum_{j,\ell,k} R_{1i(\phi_j\phi_\ell\phi_k)}(y_i, \boldsymbol{\phi}_0 + z_m u_{1j\ell k} \boldsymbol{e}_{j\ell k}) - \sum_{j,\ell,k} R_{1i(\phi_j\phi_\ell\phi_k)}(y_i, \boldsymbol{\phi}_0 + z_m u_{2j\ell k} \boldsymbol{e}_{j\ell k}) \Big|, \end{aligned}$$

for some $u_{2j\ell k} \in [-1, 1]$. Using property of (8.22), we conclude that the above term is bounded above by $C(y_i) z_m^3$, which completes the proof.

Acknowledgment

I thank my supervisor Prof. Tatsuya Kubokawa for all of his help and encouragement during the master's and doctoral courses. I also thank Prof. Satoshi Yamashita, Dr. Hisashi Noma and Dr. Takahiro Otani for their encouragement while I was working as a research fellow at the Institute of Statistical Mathematics. Prof. Alan Welsh, Prof. Stephen Hassett, Prof. Gauri Datta, Prof. Malay Ghosh, Prof. J.N.K. Rao, Prof. Naoto Kunitomo, Prof. Yoshihiro Yajima, Prof. Yasuhiro Omori, Prof. Katsumi Shimotsu, Prof. Akimichi Takemura, Prof. Fumiyasu Komaki, Prof. Yuzo Maruyama and Dr. Kengo Kato gave me many helpful comments about my research. I was grateful for valuable discussions with Dr. Masayo Hirose, Dr. Kota Ogasawara, Dr. Genya Kobayashi, Dr. Yuki Kawakubo and Mr. Hiromasa Tamae, which made my research more meaningful. I was financially supported by JSPS KAKENHI Grant Numbers JP15J10076 and JP16H07406. I would like to thank my family for their constant support throughout my life. Finally, my deepest thanks to my wife Naho for her love, support and understanding.

Bibliography

Arima, S., Datta, G. S. and Liseo, B. (2015). Bayesian estimators for small area models when auxiliary information is measured with error. *Scandinavian Journal of Statistics*, **42**, 518-529.

Arora, V. and Lahiri, P. (1997). On the superiority of the Bayesian method over the BLUP in small area estimation problems. *Statistica Sinica*, **7**, 1053-1063.

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. *Journal of the American Statistical Association*, **83**, 28-36.

Benavent, R. and Morales, D. (2016). Multivariate Fay-Herriot models for small area estimation. *Computational Statistics & Data Analysis*, **94**, 372-390.

Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd Edition. Springer, New York.

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

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

Brent, R. (1973). *Algorithms for Minimization without Derivatives.* Englewood Cliffs N.J. Prentice-Hall.

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

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

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

Dass, S. C., Maiti, T., Ren, H. and Sinha, S. (2012). Confidence interval estimation of small area parameters shrinking both means and variances. *Survey Methodology*, **38**, 173-187.

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

Datta, G.S., Hall, P. and Mandal, A. (2011b). Model selection by testing for the presence of small-area effects, and application to area-level data. *Journal of the American Statistical Association*, **106**, 362–374.

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.

Datta, G. S. and Mandal, A. (2015). Small area estimation with uncertain random effects. *Journal of the American Statistical Association*, **110**, 1735-1744.

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.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum Likelihood From Incomplete Data via the EM Algorithm (with discussion). *Journal of the Royal Statistical Society: Series B*, **39**, 1-38.

Fay, R. E. and Herriot, R. A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, **74**, 269-277.

Foster, J., Greer, J. and Thorbecke, E. (1984). A class of decomposable poverty measures. *Econometrica*, **52**, 761-766.

Godambe, V. P. and Thompson, M. E. (1989). An extension of quasi-likelihood estimation (with Discussion). *Journal of Statistical Planning and Inference*, **22**, 137-152.

Ghosh, M. and Maiti, T. (2004). Small-area estimation based on natural exponential family quadratic variance function models and survey weights. *Biometrika*, **91**, 95-112.

Ghosh, M, Natarajan, K., Stroud, T. W. F. and Carlin, B. P. (1998). Generalized linear models for small area estimation. *Journal of the American Statistical Association*, **93**, 273-282.

Hall, P. and Carroll, R.J. (1989). Variance function estimation in regression: The effect of estimating the mean. *Journal of Royal Statistical Society: Series B*, **B 51**, 3-14.

Hall, P. and Maiti, T. (2006a). On parametric bootstrap methods for small area prediction. *Journal of Royal Statistical Society: Series B*, **68**, 221-238.

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

Hwang, J. T. G., Qiu, J. and Zhao, Z. (2009). Empirical Bayes confidence intervals shrinking both mean and variances. *Journal of Royal Statistical Society: Series B*, **71**, 265-285.

Jiang, J. (1996). REML estimation: asymptotic behavior and related topics. *The Annals of Statistics*, **24**, 255-286.

Jiang, J. (2006). *Linear and generalized linear mixed models and their applications*, Springer.

Jiang, J. and Nguyen, T. (2012). Small area estimation via heteroscedastic nested-error regression. *Canadian Journal of Statistics*, **40**, 588-603.

Jones, M. C. and Pewsey, A. (2009). Sinh-arcsinh distributions. *Biometrika*, **96**, 761-780.

Kubokawa, T., Hasukawa, M. and Takahashi, K. (2014). On measuring uncertainty of benchmarked predictors with application to disease risk estimate. *Scandinavian Journal of Statistics*, **41**, 394-413.

Kubokawa, T., Sugasawa, S., Ghosh, M. and Chaudhuri, S. (2016). Prediction in heteroscedastic nested error regression models with random dispersions. *Statistica Sinica*, **26**, 465-492.

Lahiri, A. N., Maiti, T., Katzoff, M and Parsons, V. (2007). Resampling-based empirical prediction: an application to small area estimation. *Biometrika*, **94**, 469-485.

Lahiri, P. and Rao, J. N. K. (1995). Robust estimation of mean squared error of small area estimators. *Journal of the American Statistical Association*, **90**, 758-766.

Li, H. and Lahiri, P. (2010). An adjusted maximum likelihood method for solving small area estimation problems. *Journal of Multivariate Analysis*, **101**, 882-892.

Lohr, S. L. and Rao, J. N. K. (2009). Jackknife estimation of mean squared error of small area predictors in nonlinear mixed models. *Biometrika*, **96**, 457-468.

Maiti, T., Ren, H. and Sinha, A. (2014). Prediction error of small area predictors shrinking both means and variances. *Scandinavian Journal of Statistics*, **41**, 775-790.

Maples, J., Bell, W. and Huang, E. (2009). Small area variance modeling with application to county poverty estimates from the American community survey. In *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 5056-5067.

Molina, I., Nandram, B. and Rao, J. N. K. (2014). Small area estimation of general parameters with application to poverty indicators: A hierarchical Bayes approach. *The Annals of Applied Statistics*, **8**, 852-885.

Molina, I. and Marhuenda, Y. (2015). sae: an R package for small area estimation. *The R Journal*, **7**, 81-98.

Molina, I. and Rao, J. N. K. (2010). Small area estimation of poverty indicators. *Canadian Journal of Statistics*, **38**, 369-385.

Morris, C. (1982). Natural exponential families with quadratic variance functions. *The Annals of Statistics*, **10**, 65-80.

Morris, C. (1983). Natural exponential families with quadratic variance functions: Statistical theory. *The Annals of Statistics*, **11**, 515-529.

Muller, H.G. and Stadtmuller, U. (1987). Estimation of heteroscedasticity in regression analysis. *The Annals of Statistics*, **15**, 610-625.

Opsomer, J. D., Claeskens, G., Ranalli, M. G., Kauermann, G. and Breidt, F. J. (2008). Non-parametric small area estimation using penalized spline regression. *Journal of Royal Statistical Society: Series B*, **70**, 265-286.

Pfeffermann, D. (2013). New Important Developments in Small Area Estimation. *Statistical Science*, **28**, 40-68 .

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

Rao, J.N.K. and Molina, I. (2015) *Small Area Estimation, 2nd Edition*. Wiley.

Ruppert, D., Wand, M.P, Holst, U., and Hossjer, O. (1997). Local polynomial variance-function estimation. *Technometrics* **39**, 262-273.

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

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, B.*, **64**, 583-639.

Sugasawa, S. and Kubokawa, T. (2015). Parametric transformed Fay-Herriot model for small area estimation. *Journal of Multivariate Analysis*, **139**, 295-311.

Sugasawa, S. and Kubokawa, T. (2016). On conditional prediction errors in mixed models with application to small area estimation. *Journal of Multivariate Analysis*, **148**, 18-33.

Sugasawa, S. and Kubokawa, T. (2017a). Heteroscedasetic nested error regression models with variance functions *Statistica Sinica*, **27**, 1101-1123.

Sugasawa, S. and Kubokawa, T. (2017b). Transforming response values in small area prediction. *Computational Statistics & Data Analysis*, **114**, 47-60.

Sugasawa, S. and Kubokawa, T. (2017c). Bayesian estimators in uncertain nested error regression models. *Journal of Multivariate Analysis*, **153**, 52-63.

Sugasawa, S. and Kubokawa, T. (2017d). Adaptively transformed mixed model prediction of general finite population parameters. arXiv:1705.04136.

Sugasawa, S., Tamae, H. and Kubokawa, T. (2017a). Bayesian estimators of small area models shrinking both mean and variances. *Scandinavian Journal of Statistics*, **44**, 150-167.

Sugasawa, S., Kubokawa, T. and Ogasawara, K. (2017b). Empirical uncertain Bayes methods in area-level models. *Scandinavian Journal of Statistics*, **44**, 684-706.

Torabi, M. and Rao, J.N.K. (2013). Estimation of mean squared error of model-based estimators of small area means under a nested error linear regression model. *Journal of Multivariate Analysis*. **117**, 76-87.

Wang, J. and Fuller, W. (2003). The mean squared error of small area predictors constructed with estimated error variances. *Journal of the American Statistical Association*, **98**, 716-723.

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

You, Y. and Chapman, B. (2006). Small area estimation using area level models and estimated sampling variances. *Survey Methodology*, **32**, 97-103.