Noname manuscript No.
(will be inserted by the editor)

Constrained Bayes Estimation in Small Area Models with Functional Measurement Error

Elaheh Torkashvand · Mohammad Jafari Jozani · Mahmoud Torabi

Received: date / Accepted: date

Abstract In survey sampling, policy decisions regarding allocation of resources to subgroups, called small areas, or determination of subgroups with specific properties in a population are based on reliable estimates of small area parameters. However, the information is often collected at a different scale than these subgroups. Hence, we need to estimate characteristics of subgroups based on the coarser scale data. One of the main interests in small area estimation is to produce an ensemble of small area parameters whose distribution across small areas is close to the corresponding distribution of true parameters. In this paper, we consider the unit-level nested error linear regression model which is commonly used in small area estimation. We study the case where the covariate in the model is assumed to have measurement error. To study this complex model, we propose to use constrained Bayes method to estimate the true covariate in order to build the small area Bayes predictor. We also provide some measures of performance such as sensitivity, specificity, and positive/negative predictive values for the constructed Bayes predictor. We estimate the model parameters using the method of moments and Bayesian approach to get corresponding empirical and hierarchical Bayes predictors. The performance of our proposed approach is evaluated through a simulation study and a real data application.

Keywords Constrained Bayes · Empirical Bayes · Functional Measurement Error · Small Area Estimation.

Elaheh Torkashvand
Department of Statistics, University of Manitoba, Winnipeg, MB, CANADA, R3T 2N2 E-mail: torkashe@umanitoba.ca

Mohammad Jafari Jozani
Department of Statistics, University of Manitoba, Winnipeg, MB, CANADA, R3T 2N2 E-mail: M_Jafari_Jozani@umanitoba.ca

Mahmoud Torabi
Departments of Community Health Sciences and Statistics, University of Manitoba, Winnipeg, MB, CANADA, R3E 0W3 E-mail: Mahmoud.Torabi@umanitoba.ca

1. Introduction

Sample surveys are commonly conducted to provide reliable estimates of the finite population parameters such as totals, means, counts, quantiles, etc. In recent years, there has been increasing demand to get such estimates for sub-populations (small areas), such as counties or gender-age groups, due to their growing use in formulating policies and programs, allocating government funds, regional planning, marketing decisions at local level, and other uses. However, sample sizes within areas are often too small to warrant the use of the traditional area-specific direct estimates.

To produce reliable estimates of characteristics of interest for small areas and obtain measures of error associated with each estimate, different methods have been proposed in the literature. These include, among others, the use of synthetic, composite and/or model-based estimators (Datta et al, 2011; Jiang and Lahiri, 2006; Jiang, 2010; Rueda et al, 2010; Rao and Molina, 2015). Model-based estimators borrow strength from related areas by defining a set of assumptions to model the stochastic behaviour of the variables in the underlying population and by introducing random effects into the model. In the context of linear mixed models, such small area models may be classified into two broad types: (i) Area-level models that relate small area direct estimates to area-specific covariates; such models are used if unit-level data are not available. (ii) Unit-level models that relate the unit values of a study variable to associated unit-level covariates with known area means and area-specific covariates. A comprehensive account of model-based small area estimation under area-level and unit-level models is given by Rao and Molina (2015). In this paper, we focus on an empirical Bayes estimation of small area means under unit-level nested error linear regression model with measurement errors in the covariate values.

Battese et al (1988) and Prasad and Rao (1990) used a unit-level nested error linear regression model where the covariates are not subject to measurement errors. However, there are many circumstances where the covariates are subject to measurement error. In a pioneering paper, Ghosh and Sinha (2007), henceforth abbreviated GS, proposed a nested error linear regression model with an area-level covariate subject to measurement error. The Bayes predictors of small area means were obtained and consequently, the pseudo-Bayes (PB) and pseudo empirical Bayes (PEB) predictors were constructed by using the method of moments (MM) estimator of the true area-specific covariate and estimates of the model parameters. Analytically and also using simulation studies, GS showed the superiority of their method (in terms of the convergence of the PEB predictors of small area means to PB predictors) over the naive predictors of small area means that are obtained by simply neglecting the measurement error. Datta et al (2010) proposed a new PB predictor of the small area mean by using the maximum likelihood estimate (MLE) of the true area-specific covariate. Their proposed PEB predictor improved the one proposed by GS in terms of the relative bias of the estimator of the mean

squared prediction error (MSPE). Torabi (2011) extended the model given by GS using survey weights of the response values.

Torkashvand et al (2015) proposed another variant of the PB predictor of the small area mean using the James-Stein (JS) estimate of the area-level covariate. They also showed that their new PEB predictor outperforms previously proposed predictors using the MM and the MLE in terms of the MSPE and relative bias.

A drawback of the proposed JS estimator in Torkashvand et al (2015), and essentially any other method which uses the Bayesian and/or empirical Bayesian methodology to estimate the true area-specific covariate, is that the empirical histogram of these model-based predictors is underdispersed as an estimate of the histogram of the true small area means; noting that the corresponding predictors overshrink the direct estimates to their regression estimates (Lyles et al, 2007; Spjøtvoll and Thomsen, 1987). However, in practice, there are many situations where the interest lies in producing an ensemble of parameter estimates whose distribution is close to the distribution of area-specific parameters. For example, in a hypertension study (see Section 5), one might be interested in identifying small areas whose true mean (diastolic) blood pressures are either below or above certain thresholds to identify groups that are more at risk for having hypertension. Another instance is to identify small areas with average income less than the poverty line or above a specific threshold (Ghosh, 1992) in order to have a better understanding of the socio-economic status of the population.

Louis (1984) proposed constrained (empirical) Bayes estimates of small area means in order to adjust the overshrinkage of small area means towards the prior distribution for the special case of the Fay-Herriot model. Lahiri (1990) gave an exact expression for the overshrinkage of the Bayes estimates of small area means when the underlying distribution of small area means are determined only up to their mean and variance and the conditional distribution of the response variable given the small area mean belongs to an exponential family with a quadratic variance function. Ghosh (1992) found a general expression for the constrained Bayes and showed the superiority of constrained Bayes estimates of small area means over the ML estimates in terms of the Bayes risk for the normal distribution.

Since Louis (1984), constrained Bayes (CB) estimation has been widely used in small area estimation for different purposes. For example, Datta et al (2011) used the CB approach for Bayesian benchmarking in small area estimation to provide an overall agreement between model-based area estimates and direct estimates at an aggregate level. Later Ghosh and Steorts (2013) extended these results to a multi-stage benchmarking scenario. Ha (2013) proposed a general benchmarking method for complex benchmarking questions. Later on, Ha and Lahiri (2014) pointed out potential problems caused by benchmarking and recommended to implement the benchmarking with caution. See Jiang and Lahiri (2006), Ugarte et al (2009),Kubokawa and Strawderman (2013), Pfeffermann et al (2014) and references therein, for other references on the

theory and applications of the CB method in different problems including small area estimation.

Our goal in this paper is to implement the CB methodology in small area estimation problems with measurement errors when the interest lies in getting a more precise picture of the true structure of small area parameters in order to make classifications rather than point estimation of individual area parameters. To this end, we propose to use the CB method to adjust the estimators of true area-specific covariates and consequently construct new PB predictors of small area means. We show that using this approach, one can obtain a more accurate estimate of the underlying histogram of the true area-specific covariate subject to the functional measurement error, and consequently a more precise histogram of small area predictors. Constructed CB predictors can be used for several purposes, including ranking among areas, detection of extremes, etc.

The outline of the paper is as follows. In Section 2, we study the unit-level regression models with the functional measurement error in the area-specific covariate and construct the CB estimator of the true area-specific covariate. Further, we obtain the PB predictors of area means based on the CB estimate of the true area-specific covariate which dominate the PB predictors of area means based on the ML estimate of the true area-specific covariate. In Sections 3 and 4, we obtain the constrained empirical Bayes (CEB) and constrained hierarchical Bayes (CHB) predictors of small area means and evaluate the performance of different predictors using some statistical criteria. The performance of our proposed approach is evaluated using a real data (a blood pressure study in New Zealand) and a simulation study in Sections 5 and 6, respectively. Some concluding remarks are given in Section 7.

2. Constrained Bayes Estimates of the True Area-Specific Covariate

We consider the following nested error linear regression population model

$$y_{ij} = b_0 + b_1 x_i + u_i + e_{ij}, \quad i = 1, \ldots, m, j = 1, \ldots, N_i, \tag{2.1}$$

with

$$X_{ij} = x_i + \eta_{ij}, \quad i = 1, \ldots, m, j = 1, \ldots, N_i, \tag{2.2}$$

where N_i is the known population size of the ith area (i=1,...,m), y_{ij} is the value of the study variable associated with the jth unit in the ith area and x_i is the unknown true area-specific covariate associated with y_{ij}. Further, the random errors e_{ij}, measurement errors η_{ij} and the area-level random effects u_i are assumed to be mutually independent with $e_{ij} \overset{i.i.d}{\sim} N(0, \sigma_e^2)$, $\eta_{ij} \overset{i.i.d}{\sim} N(0, \sigma_\eta^2)$ and $u_i \overset{i.i.d}{\sim} N(0, \sigma_u^2)$. We assume that x_i is a fixed and unknown parameter which is called functional measurement error (Fuller, 2009). When there is no measurement error in covariate (i.e., $\sigma_\eta^2 = 0$) we have $x_i = X_{ij}$, and (2.1) and (2.2) reduce to the unit level regression model

$$y_{ij} = b_0 + b_1 X_{ij} + u_i + e_{ij}, \quad i = 1, \ldots, m, j = 1, \ldots, N_i. \tag{2.3}$$

proposed by Battese et al (1988). Assuming that there is no sample selection bias (i.e. the sampling scheme is not informative), a sample of size n_i is selected from the i'th area and the sample data, without loss of generality, is denoted by $(\mathbf{y}, \mathbf{X}) = \{(y_{ij}, X_{ij}),\ i = 1, \ldots, m,\ j = 1, \ldots, n_i\}$. Also, it is assumed that the covariate is only observed for the units in the sample.

Our main interest is to estimate small area means, $\gamma_i = \frac{1}{N_i}\sum_{j=1}^{N_i} y_{ij}$, for $i = 1, \ldots, m$. Ghosh and Sinha (2007) obtained the PB predictor of γ_i as

$$\widehat{\gamma}_i^{PB} = \widehat{\gamma}_i^{PB}(x_i, \phi) = (1 - f_i B_i)\bar{y}_i + f_i B_i(b_0 + b_1 x_i), \tag{2.4}$$

where $\phi = (b_0, b_1, \sigma_u^2, \sigma_e^2)$, $f_i = 1 - \frac{n_i}{N_i}$ is the finite population correction factor, and $B_i = \frac{\sigma_e^2}{\sigma_e^2 + n_i\sigma_u^2}$, for $i = 1, \ldots, m$. However, (2.4) depends on x_i's, the true values of the small area-specific covariate that are unknown. We need to obtain estimates of x_i's. Estimators of the true area-specific covariate are proposed under the quadratic loss function, $L(x_i, \lambda(\mathbf{y}, \mathbf{X})) = (x_i - \lambda(\mathbf{y}, \mathbf{X}))^2$, where $\lambda(\mathbf{y}, \mathbf{X})$ is the estimator of the true area-specific covariate, such that some optimal properties are preserved. Carroll et al (2010)[Sec. 9.1.3] and Carroll et al (1999) considered normal prior distribution for the functional measurement error. Torkashvand et al (2015) used the same idea to derive the Bayes estimator of the true area-specific covariate when $x_i \overset{i.i.d.}{\sim} N(\mu, \tau^2)$ as

$$x_{iB}(\bar{Z}_i^*) = \mathrm{E}(x_i|\bar{Z}_i^*) = C_i\mu + (1 - C_i)\bar{Z}_i^*, \quad i = 1, \ldots, m, \tag{2.5}$$

where $C_i = \frac{\sigma_i^2}{\sigma_i^2 + \tau^2}$, $\sigma_i^2 = \frac{\sigma_\eta^2\left(\sigma_u^2 + \frac{\sigma_e^2}{n_i}\right)}{n_i\sigma_u^2 + \sigma_e^2 + b_1^2\sigma_\eta^2}$, and $\bar{Z}_i^* = \bar{X}_i + \frac{b_1\sigma_\eta^2}{\sigma_e^2 + n_i\sigma_u^2 + b_1^2\sigma_\eta^2}(\bar{y}_i - b_0 - b_1\bar{X}_i)$ is the ML estimator of the true area-specific covariate (Datta et al, 2010), $\bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$ and $\bar{X}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}$, for $i = 1, \ldots, m$.

Even though the Bayesian method gives the best estimate of x_i's in terms of the minimum Bayes risk, it overshrinks the Bayes estimates toward the prior mean (μ) in the sense that

$$\mathrm{E}\Big(\sum_{i=1}^{m}(x_i - \bar{x})^2|\bar{\mathbf{Z}}^*\Big) \geq \sum_{i=1}^{m}(x_{iB}(\bar{Z}_i^*) - \bar{x}_B(\bar{\mathbf{Z}}^*))^2, \tag{2.6}$$

where $\bar{x}_B(\bar{\mathbf{Z}}^*) = \frac{1}{m}\sum_{i=1}^{m} x_{iB}(\bar{Z}_i^*)$, $\bar{\mathbf{Z}}^* = (\bar{Z}_1^*, \bar{Z}_2^*, \ldots, \bar{Z}_m^*)$, and $\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$. Formula (2.6) states that the variance of x_{iB} is less than the posterior variance of x_i's. In other words, as x_{iB}'s are marginally unbiased estimators of μ, the inequality implies that they are more concentrated around μ in comparison with the true posterior distribution of x_i's. The equality holds if and only if all $(x_1 - \bar{x}), \ldots, (x_m - \bar{x})$ have degenerate posterior distributions (Ghosh, 1992). This overshrinkage results in the poor performance of the Bayes estimate in tails of the histogram of x_i's (Ghosh, 1992; Louis, 1984; Lyles et al, 2007), and consequently, the resulting PB predictors of small area means based on the Bayes estimates of x_i's will perform poorly (see Sections 6 and 7). Louis (1984) proposed to use the CB estimation method in order to address the problem of overshrinkage of the standard Bayes estimator toward the prior

mean, when the underlying distribution is a multivariate normal distribution. This was later extended by Ghosh (1992) to other distributions. Following Ghosh (1992), the CB estimator of x_i is obtained by minimizing the posterior risk under the sum of squared error loss function

$$\mathrm{E}\left[\sum_{i=1}^{m}(x_i - t_i)^2|\bar{\mathbf{Z}}^*\right],$$

within the class of all estimators $\mathbf{t}(\bar{\mathbf{Z}}^*) = \mathbf{t} = (t_1, \ldots, t_m)$ of $x = (x_1, \ldots, x_m)$ that satisfy the following two conditions

$$\mathrm{E}(\bar{x}|\bar{\mathbf{Z}}^*) = \tfrac{1}{m}\sum_{i=1}^{m} t_i(\bar{\mathbf{Z}}^*) = \bar{t}(\bar{\mathbf{Z}}^*),$$
$$\mathrm{E}(\sum_{i=1}^{m}(x_i - \bar{x})^2|\bar{\mathbf{Z}}^*) = \sum_{i=1}^{m}(t_i(\bar{\mathbf{Z}}^*) - \bar{t}(\bar{\mathbf{Z}}^*))^2.$$

These conditions will help to obtain modified Bayes estimates of the values of the true area-specific covariate that have a histogram with the same mean and spread as the posterior mean and variance of the histogram of the true area-specific covariate. Now, using the Lagrange's method of undetermined multipliers, we obtain the CB estimators of x_i's as follow

$$x_{iCB}(\bar{\mathbf{Z}}^*) = \nu x_{iB}(\bar{Z}_i^*) + (1-\nu)\bar{x}_B(\bar{\mathbf{Z}}^*), \tag{2.7}$$

where

$$\nu \equiv \nu(\bar{\mathbf{Z}}^*) = \left(1 + \frac{H_1(\bar{\mathbf{Z}}^*)}{H_2(\bar{\mathbf{Z}}^*)}\right)^{\frac{1}{2}}, \tag{2.8}$$

and

$$H_1(\bar{\mathbf{Z}}^*) = (1-\tfrac{1}{m})\sum_{i=1}^{m}\frac{\tau^2\sigma_i^2}{\tau^2+\sigma_i^2} = (1-\tfrac{1}{m})\sum_{i=1}^{m}\tau^2 C_i$$
$$H_2(\bar{\mathbf{Z}}^*) = \sum_{i=1}^{m}(x_{iB}(\bar{\mathbf{Z}}^*) - \bar{x}_B(\bar{\mathbf{Z}}^*))^2. \tag{2.9}$$

Due to (2.8), ν has the stochastic nature. In Lemma 1, we present an almost sure asymptotic value of ν, say $\widehat{\nu}$. Using $\widehat{\nu}$, we show the optimality of CB estimators of x_i's over their corresponding ML estimators in Theorem 1.

Lemma 1 *Suppose the model parameters are known. As $m \to \infty$, ν defined in (2.8), almost surely converges to*

$$\widehat{\nu} = \left(1 + \frac{(1-\frac{1}{m})\sum_{i=1}^{m} C_i}{\sum_{i=1}^{m}(1-C_i)}\right)^{\frac{1}{2}}. \tag{2.10}$$

Proof See the supplementary document for the proof.
We keep using the asymptotic value of ν defined in (2.10) throughout the paper due to the simplicity it introduces into the analysis. In Theorem 1, the aim is to show the superiority of the CB estimate of x_i over the ML estimate in terms of the Bayes risk when we plug in $\widehat{\nu}$ in (2.7) as an estimate of ν. To this end, the Bayes risk is defined as

$$r(\pi, \lambda(\mathbf{y}, \mathbf{X})) = \mathrm{E}_{\mathbf{x}} \left[\mathrm{E}_{(\mathbf{y},\mathbf{X})} \left(L(x_i, \lambda(\mathbf{y}, \mathbf{X})) \right) \right].$$

Theorem 1 *Consider the models (2.1) and (2.2) with known model parameters. Suppose the prior distribution on* $\mathbf{x} = (x_1, \ldots, x_m)$ *is* $\pi \sim N(\mu, \tau^2)$ *and the squared error loss function is used as the underlying loss function. Using the asymptotic value of* ν *defined in Lemma 1, the CB estimators,* $\mathbf{x}_{CB} = (x_{1CB}, \ldots, x_{mCB})$*, of true area-specific covariates,* $\mathbf{x}$*, dominate the corresponding ML estimators,* $\overline{\mathbf{Z}}^* = (\bar{Z}_1^*, \bar{Z}_2^*, \ldots, \bar{Z}_m^*)$*, in terms of the Bayes risk, that is* $r(\pi, \mathbf{x}_{CB}) < r(\pi, \overline{\mathbf{Z}}^*)$.

Proof See the supplementary document for the proof.
Ghosh and Sinha (2007) introduced the PB predictor of γ_i as

$$\widehat{\gamma}_{iGS}^{PB} = \widehat{\gamma}_i^{PB}(x_{iGS}, \phi) = (1 - f_i B_i)\bar{y}_i + f_i B_i(b_0 + b_1 x_{iGS}), \tag{2.11}$$

where $x_{iGS} = \bar{X}_i$ for $i = 1, \ldots, m$. Similarly, Datta et al (2010) introduced the PB predictor of small area means based on the ML estimate of the true area-specific covariate, $x_{iML} = \bar{Z}_i^*$, for $i = 1, \ldots, m$. In Torkashvand et al (2015), the James-Stein estimate of the true area-specific covariate, x_{iJS}, was used to construct a new PB predictor of γ_i. In this paper, we introduce a new PB predictor of small area means based on the CB estimate of the true area-specific covariate by replacing x_{iGS} with x_{iCB} in (2.11), i.e.

$$\widehat{\gamma}_{iCB}^{PB} = \widehat{\gamma}_i^{PB}(x_{iCB}, \phi) = (1 - f_i B_i)\bar{y}_i + f_i B_i(b_0 + b_1 x_{iCB}). \tag{2.12}$$

In Theorem 2, we also show that $\widehat{\gamma}_{iCB}^{PB}$ dominates $\widehat{\gamma}_{iML}^{PB}$ in terms of the Bayes risk.

Theorem 2 *The PB predictor of small area mean based on the CB estimate of the true area-specific covariate dominates the PB predictor of small area mean based on the ML estimate of the true area-specific covariate in terms of the Bayes risk*

$$\sum_{i=1}^{m} r(\pi, \widehat{\gamma}_{iCB}^{PB}) \leq \sum_{i=1}^{m} r(\pi, \widehat{\gamma}_{iML}^{PB}).$$

Proof See the supplementary document for the proof.
Similar to (2.11), $\widehat{\gamma}_{iCB}^{PB}$ depends on the unknown model parameters. In Section 4, we provide estimates of the parameters using the method of moments, empirical Bayes, and hierarchical Bayes methods. To compare the performance of these predictors, we use different measures of performance such as the sensitivity (Se), specificity (Sp), positive predictive value (PPV), and negative

predictive value (NPV). These measures provide some necessary probabilities
215 that are useful to evaluate the precision of the PB predictors of the small area means when the goal is to study whether the small area means are above or under a specified threshold (Lyles and Xu, 1999).

3. Performance measures

In Section 2, we introduced the CB estimator of the true area-specific covariate. Some optimal properties of the CB estimator of the true area-specific covariate were also discussed in the Bayesian set-up. In practice, researchers sometimes want to classify areas according to whether their means are above or below some meaningful thresholds. For example, in the diastolic blood pressure application, having a blood pressure above the threshold $t = 80$ indicates the pre-hypertension phase (Zhang and Li, 2011). In Sections 5 and 6, we use the proposed estimator to study such scenarios. To this end, small area means are predicted using different approaches. People inside each area are considered to be in a pre-hypertension phase if their area mean is greater than 80. To evaluate the performance of the proposed estimators, we use the criteria defined in Lyles and Xu (1999). Consider t as a disease diagnostic threshold. In this work, we assume that a "positive" test or suffering from a "disease" happens if "$\widehat{\gamma}_i^{PB} > t$" where we treat $\widehat{\gamma}_i^{PB}$ as the predictor of the small area mean and a diagnostic test. If being below the threshold indicates of "disease" and "positive" test, some adjustments of these definitions are required. Following Lyles and Xu (1999), some statistical properties of the candidate predictors are specified as

$$
\begin{aligned}
Se &= \mathbb{P}(\text{“Positive” given “disease”}) = \mathbb{P}(\widehat{\gamma}_i^{PB} > t | \gamma_i > t),\\
Sp &= \mathbb{P}(\text{“Negative” given “no disease”}) = \mathbb{P}(\widehat{\gamma}_i^{PB} < t | \gamma_i < t),\\
PPV &= \mathbb{P}(\text{“disease” given “Positive”}) = \mathbb{P}(\gamma_i > t | \widehat{\gamma}_i^{PB} > t),\\
NPV &= \mathbb{P}(\text{“no disease” given “Negative”}) = \mathbb{P}(\gamma_i < t | \widehat{\gamma}_i^{PB} < t).
\end{aligned}
$$

Analytical expressions for these quantities can be obtained following (2.1) and (2.2) and the bivariate normal distribution of $(\gamma_i, \widehat{\gamma}_i^{PB})$. In particular, we have

$$
\begin{aligned}
Se_{iCB} &= \frac{\mathbb{P}(\widehat{\gamma}_{iCB}^{PB} > t, \gamma_i > t)}{\mathbb{P}(\gamma_i > t)} = \frac{\mathbb{P}(S_i > (t - \alpha_i u_i), u_i > d_i)}{p}\\
&= \frac{\int_{d_i}^{\infty} \Phi(\frac{\mu_{s_i} - t + \alpha_i u_i}{\sigma_{s_i}}) f(u) du}{p}, \qquad (3.1)
\end{aligned}
$$

where $p = \mathbb{P}(\gamma_i > t)$, $d_i = t - b_0 - b_1 x_i$, and

$$
\begin{aligned}
S_i = (1 - f_i B_i)(b_0 + b_1 x_i) + f_i B_i b_1 A_i x_i + f_i B_i \sum_{j \neq i} D_j \bar{Z}_j^*\\
+ \alpha_i \bar{e}_i + F_i \bar{\eta}_i + f_i B_i \left(b_0 + b_1 \mu (\nu C_i + (1 - \nu) \tfrac{\sum_{j=1}^{m} C_j}{m}) \right).
\end{aligned}
$$

Also, $A_i = (1-C_i)(\nu + \frac{1-\nu}{m})$, $\alpha_i = 1 - f_iB_i\left(1 - A_i\frac{b_1^2\sigma_\eta^2}{\sigma_e^2+n_i\sigma_u^2+b_1^2\sigma_\eta^2}\right)$, $F_i = f_iB_ib_1A_i\left(1-\frac{b_1^2\sigma_\eta^2}{\sigma_e^2+n_i\sigma_u^2+b_1^2\sigma_\eta^2}\right)$ and $D_j = \frac{1}{m}(1-\nu)(1-C_j)$. Moreover, we have $S_i \sim N(\mu_{s_i}, \sigma_{s_i}^2)$, where $\mu_{s_i} = (1-f_iB_i)(b_0+b_1x_i) + f_iB_ib_1A_ix_i + f_iB_i(b_0 + b_1\mu(\nu C_i + (1-\nu)\frac{\sum_{j=1}^m C_j}{m})) + f_iB_i\sum_{j\neq i}D_jx_j$ and $\sigma_{s_i}^2 = \alpha_i^2\frac{\sigma_e^2}{n_i} + F_i^2\frac{\sigma_\eta^2}{n_i} + (f_iB_i)^2\sum_{j\neq i}D_j^2\sigma_j^2$. Similarly, we have

$$Sp_{iCB} = \frac{\mathbb{P}(\widehat{\gamma}_{iCB}^{PB} < t, \gamma_i < t)}{\mathbb{P}(\gamma_i < t)} = \frac{\mathbb{P}(S_i < (t-\alpha_iu_i), u_i < d_i)}{1-p}$$
$$= \frac{\int_{-\infty}^{d_i}\Phi(\frac{t-\alpha_iu_i-\mu_{s_i}}{\sigma_{s_i}})f(u)du}{1-p}. \tag{3.2}$$

Also,

$$PPV_{iCB} = \frac{pSe_{iCB}}{\Phi\left(-\frac{(t-\mathrm{E}(\widehat{\gamma}_{iCB}^{PB}))}{\sigma_{\widehat{\gamma}_{iCB}^{PB}}}\right)} \quad \text{and} \quad NPV_{iCB} = \frac{(1-p)Sp_{iCB}}{\Phi\left(\frac{(t-\mathrm{E}(\widehat{\gamma}_{iCB}^{PB}))}{\sigma_{\widehat{\gamma}_{iCB}^{PB}}}\right)}, \tag{3.3}$$

where

$$\mathrm{E}(\widehat{\gamma}_{iCB}^{PB}) = (1-f_iB_i)(b_0+b_1x_i) + f_iB_i(b_0+b_1\mathrm{E}(x_{iCB})), \tag{3.4}$$
$$var(\widehat{\gamma}_{iCB}^{PB}) = (\sigma_u^2 + \tfrac{\sigma_e^2}{n_i})[(1-f_iB_i)^2 + 2(1-f_iB_i)(f_iB_i)A_i\tfrac{b_1^2\sigma_\eta^2}{\sigma_e^2+n_i\sigma_u^2+b_1^2\sigma_\eta^2}]$$
$$+ (f_iB_ib_1)^2var(x_{iCB}), \tag{3.5}$$
$$\mathrm{E}(x_{iCB}) = \mu(\nu C_i + (1-\nu)\tfrac{\sum_{j=1}^m C_j}{m}) + A_ix_i + \sum_{j\neq i}D_jx_j, \tag{3.6}$$
$$var(x_{iCB}) = A_i^2\sigma_i^2 + \sum_{j\neq i}D_j^2\sigma_j^2, \tag{3.7}$$

and $\sigma_{\widehat{\gamma}_{iCB}^{PB}} = \sqrt{var(\widehat{\gamma}_{iCB}^{PB})}$.

As there are unknown parameters in (3.1), (3.2), and (3.3), the method of moments and the hierarchical Bayes estimates of the model parameters are used to obtain estimates of the model parameters, and consequently using the numerical method of integration, estimates of the performance measures introduced in this section are given (see Section 5 for more details).

4. The Constrained Empirical (and Hierarchical) Pseudo-Bayes Predictor of Small Area Means

In Sections 2 and 3, $\widehat{\gamma}_{iCB}^{PB}$, Se_i, Sp_i, PPV_i, and NPV_i are obtained under the assumption that model parameters are known. To estimate parameters, two scenarios are considered. First, similar to Torkashvand et al (2015), we follow Efron and Morris (1975) and Ghosh and Sinha (2007) to obtain the empirical Bayes (EB) estimates of τ^2 and μ and also the method of moments estimates of

the model parameters (GS), respectively. Consequently, the CEB of the area-level covariate and PEB predictor of small area means are given by replacing unknown parameters in equations (2.7) and (2.12).

In the second scenario, we propose the CHB estimator of the true area-specific covariate using priors on hyper parameters $\psi = (b_0, b_1, \sigma_u^2, \sigma_e^2, \sigma_\eta^2)$, μ, and τ^2. We consider informative prior distributions for the purpose of the analysis (more details are given in Section 5 and 6). We also derive Pseudo Hierarchical Bayes (PHB) predictors of small area means based on the CHB using the posterior means of the components of ψ. In Section 5 and 6, we also evaluate the performance of PEB and PHB predictors based on the EB, GS, and HB estimates of the model parameters, respectively.

Let ν_{HB} and ν_{EB} show estimates of the asymptotic value of ν using the HB, and also GS and EB estimates of the model parameters, respectively. Ghosh and Maiti (1999) pointed out that $\nu_{HB} > \nu_{EB}$ in their set-up, showing that the individual estimates shrink toward the overall average to a lesser extend by using the hierarchical Bayes method. They also found that the hierarchical and empirical Bayes methods asymptotically show the same behaviour in estimating ν. In our nested error linear regression model with the functional measurement error, we are not able to compare the behaviour of ν_{HB} and ν_{EB} mathematically due to their complicated forms. In Section 6, we evaluate their performances using simulation studies.

5. An Application

In this section, we analyze a cross-sectional data from the New Zealand population as an application of the proposed approach. This dataset is available as `xs.nz` in `VGAMdata package` in `R`. The aim is to predict the diastolic blood pressure using cholesterol level as the covariate and to determine groups of people (small areas) who are in danger of hypertension corresponding to areas with means above the specified threshold, 80, as the prehypertension phase. As the aim is to find the proportion of small areas in the upper tail of the distribution of γ_i's, it is reasonable to choose $\widehat{\gamma}_{iCEB}^{PEB}$ or $\widehat{\gamma}_{iCHB}^{PEB}$ over $\widehat{\gamma}_{iEB}^{PEB}$ and $\widehat{\gamma}_{iHB}^{PEB}$ due to the overshrinkage of either $\widehat{\gamma}_{iEB}^{PEB}$ or $\widehat{\gamma}_{iHB}^{PEB}$ towards the prior mean. Using this criterion, groups of people who are likely to be in hypertension phase are recognized. Medical treatments can be applied accordingly.

To this end, we consider the female participants of the study whose ethnicity are either Maori or others (Chinese, Indian, and other) as the population of interest. We categorize the female participants based on the age group, BMI, ethnic, and smoking status. The five number summary (the minimum, first quartile, second quartile, third quartile, and maximum) of age is obtained as 16, 32, 42, 52, and 88 while the five number summary of BMI is obtained as 12.80, 23.53, 25.86, 28.68, and 88.43. The range between two consecutive numbers of the five number summary is considered as a level for either age or BMI. The smoking status also has two levels (0 and 1). Small areas are

defined using the crossings of factors. Table 1 gives a description of some of the areas. In total, we have $m = 64$ small areas (domains) with 43 of them having the sample size n_i, $(i = 1, \ldots, 43)$, ranging from 1 to 15 and the rest having no sample. We expect the sampling error of the covariate in each small area to be negligible in comparison with the measurement error due to the grouping. Moreover, as cholesterol level and blood pressure are measured with different devices, the assumption of the independency of e_{ij}'s and η_{ij}'s holds. We consider x_i to be the true mean value of the cholesterol level in i'th area. Figure 1 shows the diastolic blood pressure versus cholesterol level for female participants with ethnicity as Maori or others. We use equations (2.1) and (2.2) to model the data, where X_{ij} is the observed value of cholesterol level and y_{ij} is the diastolic blood pressure for the j'th person in the i'th small area.

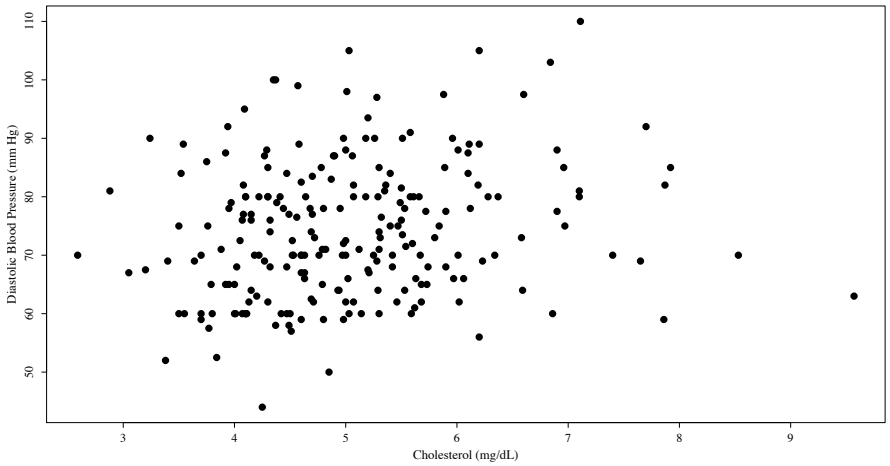

Fig. 1 Diastolic blood pressure versus cholesterol level for female participants with ethnicity as Maori or others in a cross-sectional study in New Zealand

Using the method of moments (Ghosh and Sinha, 2007), we get $\widehat{b}_{0GS} = 24.62$, $\widehat{b}_{1GS} = 9.86$, $\widehat{\sigma}^2_{eGS} = 93.39$, $\widehat{\sigma}^2_{uGS} = 26.07$, and $\widehat{\sigma}^2_{\eta GS} = 0.97$. The empirical Bayes method also results in $\widehat{\mu}_{EB} = 5.06$ and $\widehat{\tau}^2_{EB} = 0.15$. In order to apply the hierarchical Bayes method, we use the GS estimates of the model parameters and the empirical Bayes estimates of μ and τ^2 to define the prior distributions as Unif$(0, 2(9.66))$, Unif$(0, 2(5.1))$, Unif$(0, 2(0.98))$, Unif$(0, 2(0.38))$, $N(24.62, 4)$, $N(9.86, 4)$, and $N(5.063, 9)$ for σ_e, σ_u, σ_η, τ, b_0, b_1, and μ, respectively. The hierarchical Bayes method results in $\widehat{b}_{0HB} = 24.50$, $\widehat{b}_{1HB} = 9.93$, $\widehat{\sigma}^2_{eHB} = 97.60$, $\widehat{\sigma}^2_{uHB} = 2.77$, $\widehat{\sigma}^2_{\eta HB} = 1.03$, $\widehat{\mu}_{HB} = 5.04$ and $\widehat{\tau}^2_{HB} = 0.46$. It is worth mentioning that small values of $\widehat{\sigma}^2_\eta$ (either $\widehat{\sigma}^2_{\eta GS}$ or $\widehat{\sigma}^2_{\eta HB}$) refer to the variance of X_{ij}'s (ranging between 2 and 10) while $\widehat{\sigma}^2_e$ (either $\widehat{\sigma}^2_{eGS}$

Table 1 The description of some areas that are obtained using age and BMI quantiles for the purpose of grouping. The ethnicity refers to either Maori or other (Chinese, Indian, and other).

Area	Age	BMI	Ethnic	Smoking Status	n_i
1	16-31	12.80-23.52	Maori	No	13
2	16-31	23.53-25.85	Maori	No	8
5	16-31	12.80-23.52	Other	No	15
6	16-31	23.53-25.85	Other	No	1
7	16-31	25.86-28.67	Other	No	0
61	52-88	12.80-23.52	Other	Yes	1
62	52-88	23.53-25.85	Other	Yes	0

or $\widehat{\sigma}^2_{eHB}$) refers to the variance of Y_{ij}'s (ranging from 40 to 100). Further, we get $\nu_{HB} = 1.28$ and $\nu_{EB} = 1.47$. Figure 2 presents the histograms of the average cholesterol level, x_i, using different approaches. As Figure 2 shows, the histograms of the CEB and CHB estimates of x_i's have larger variance than the variance of the EB and HB estimates of the x_i's. As we show in the simulation study in Section 6, the CEB and CHB result in ensemble estimates of the true area-specific covariate with the histograms being close to the true histogram of x_i's in comparison with the EB and HB methods. The estimated values of diastolic blood pressure means, $\widehat{\gamma}_i$'s, using different approaches are given in Figure 3. We use $\widehat{\gamma}_i$'s to determine areas that are in danger of hypertension. Finally, Se, Sp, PPV, and NPV values of different predictors of small area means are given in Figure 4. In order to obtain these values, we used `integrate` function in `R` to approximate the integrals. Sometimes, the obtained values of the measures of performance are greater than one or less than zero due to the approximation so, we adjust them by projecting those values into $[0, 1]$.

As being in pre-hypertension phase is equivalent to having the diastolic blood pressure greater than 80 (Zhang and Li, 2011), Se and PPV are more important measures of performance for the current application. Figure 4 shows the pseudo estimates of Se, Sp, PPV, and NPV based on $\widehat{\gamma}^{PEB}_{iML}$, $\widehat{\gamma}^{PEB}_{iCEB}$, $\widehat{\gamma}^{PEB}_{iEB}$, $\widehat{\gamma}^{PEB}_{iCHB}$, and $\widehat{\gamma}^{PEB}_{iHB}$. We observe that $\widehat{\gamma}^{PEB}_{iCEB}$ shows the best performance in terms of the Se and PPV. But, in terms of the Sp and NPV, they have the worst performance. In the case of areas with no sample units, the ML estimate of the true area-specific covariate cannot be defined, and we have $\widehat{\gamma}^{PEB}_{iML} = \widehat{b}_0$. However, in the EB, CEB, HB, and CHB methods, to estimate the true area-specific covariate, we use the information from other areas as well. As (2.5) shows, for areas with no sample units we have $C_i = 1$ and so $x_{iB} = \mu$. Consequently, x_{iCB} can be defined using formula (2.7) for small areas with no sample units. The EB, CEB, HB and CHB estimates of the true area specific covariate are calculated by replacing the EB and GS, and also HB estimates of model parameters, respectively. The PEB predictors of small area means can be found accordingly. In the application, there are 21 small areas with no sampled units. The predicted values of small area means for areas with no sample units are $\widehat{\gamma}^{PEB}_{iML} = 24.62$,

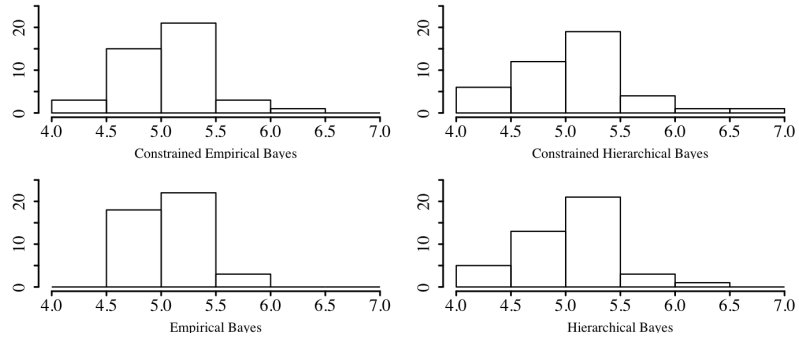

Fig. 2 Average Cholesterol Level Estimates Using Different Approaches

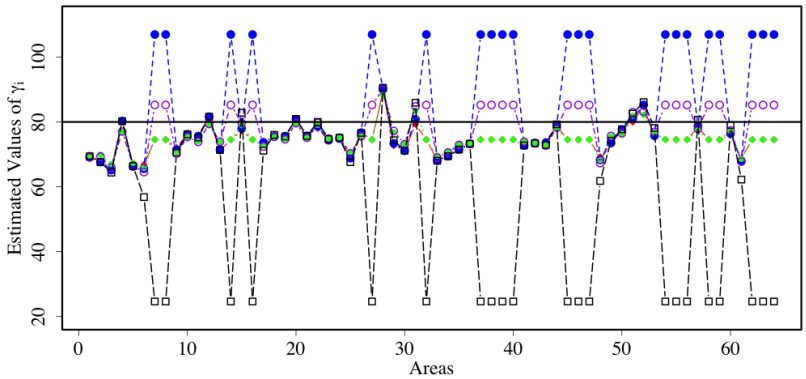

Fig. 3 The estimated values of the small area mean: □, ■, ●, ●, and ○ are corresponding to $\widehat{\gamma}_{ML}^{PEB}$, $\widehat{\gamma}_{EB}^{PEB}$, $\widehat{\gamma}_{CEB}^{PEB}$, $\widehat{\gamma}_{HB}^{PEB}$, and $\widehat{\gamma}_{CHB}^{PEB}$

$\widehat{\gamma}_{iEB}^{PEB} = 74.54$, $\widehat{\gamma}_{iHB}^{PEB} = 74.54$, $\widehat{\gamma}_{iCEB}^{PEB} = 106.97$, and $\widehat{\gamma}_{iCHB}^{PEB} = 85.23$. As detecting people who suffer from hypertension is important and also the simulation studies indicate that the CEB shows the best performance in terms of the Se and PPV, we recommend to use $\widehat{\gamma}_{iCEB}^{PEB}$ to obtain estimates of small area means.

Based on our results, areas with high diastolic blood pressure belong to overweight women. Based on our dataset, we conclude that the smoking status dose not have a significant contribution on the diastolic blood pressure. Our analysis also detects the age as an influential factor as old women with high BMI have high diastolic blood pressure.

6. Simulation Study

In this section, we implement a simulation study to evaluate the performance of the proposed method. The estimates of model parameters obtained from Section 5 are used for the purpose of simulation studies. To this end, we

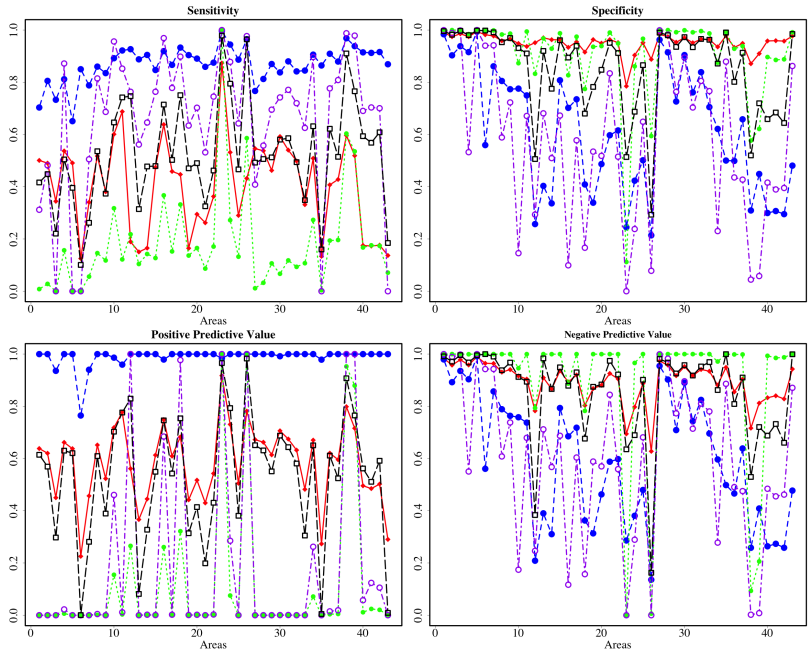

Fig. 4 The estimated values of the measures of performance of different areas: □▬, ■▬, ●╌, ●···, and ○╌ are corresponding to $\widehat{\gamma}_{ML}^{PEB}$, $\widehat{\gamma}_{EB}^{PEB}$, $\widehat{\gamma}_{CEB}^{PEB}$, $\widehat{\gamma}_{HB}^{PEB}$, and $\widehat{\gamma}_{CHB}^{PEB}$

generate $R = 5000$ simulations with $m = 43$ areas in each simulation and $\{(y_{ij}^{(r)}, X_{ij}^{(r)}); i = 1, \ldots, 43, j = 1, \ldots, n_i, r = 1, \ldots, 5000\}$ where $b_0 = 24.62$, $b_1 = 9.86$, $\sigma_e^2 = 93.39$, $\sigma_u^2 = 26.07$, $\sigma_\eta^2 = 0.97$, $\mathbf{x} = (4.47, 4.8, 4.47, 4.83, 4.31, 4.54,4.64, 5.01, 4.85, 5.18, 5.34, 5.35, 4.91, 5.09, 4.95, 5.33, 5.03, 5.45, 5.08, 5.10, 4.88, 5.04, 6.34, 5.54, 5.07, 5.84, 4.67, 4.83, 5.05, 4.93, 5.11, 4.93, 4.86, 5.25, 4.68, 5.26,5.08, 5.80, 5.49, 5.20, 5.18, 5.22, 4.72)$ where x's are the estimates of true area specific cholesterol level obtained in Section 5, and the sample size $n = (13, 8, 5, 10, 15, 1, 4, 7, 4, 9, 12, 1, 1, 1, 6, 9, 5, 3, 1, 2, 2, 3, 10, 4, 2, 2, 13, 10, 5, 12, 7, 7, 3, 5, 1, 3, 4, 4, 4, 1, 1, 1, 1)$. Further, we set $N_i = 100n_i$ for $i = 1, \ldots, m$.

In Section 5, we obtained different estimates of the model parameters using GS and HB methods. Table 2 gives the mean squared error (MSE) of the model parameters. Based on the MSE, it seems that we get more reliable estimates of σ_u^2 and σ_e^2 using the GS method. Our result (not shown here) indicates that using the GS method, τ^2 is estimated close to the true one while the HB method overestimates τ^2 significantly.

As the stochastic nature of ν introduces many difficulties in explaining the optimal properties of the CB estimator of the true area-specific covariate, we obtained an asymptotic expression for it in Lemma 1. In this section, we evaluate the precision of this asymptotic value. Figure 5 shows the asymptotic values of ν versus its exact values using the hierarchical and empirical Bayes

Table 2 The MSE of the model parameters using GS and HB methods

	b_0	b_1	σ_e^2	σ_u^2	σ_η^2
MSE GS	42147.00	1647.23	101.49	214.07	0.01
MSE HB	42040.84	1643.08	217.07	507.76	0.02

estimates of the model parameters. We observe that the empirical Bayes estimates of the model parameters perform better than the corresponding HB estimates in terms of the asymptotic behaviour of ν. It is worth mentioning that since (2.10) is a decreasing function of τ^2 and the HB method tends to overestimate τ^2 we expect to see $\nu_{HB} \leq \nu_{EB}$. This is also confirmed in our simulation study presented in Figure 5.

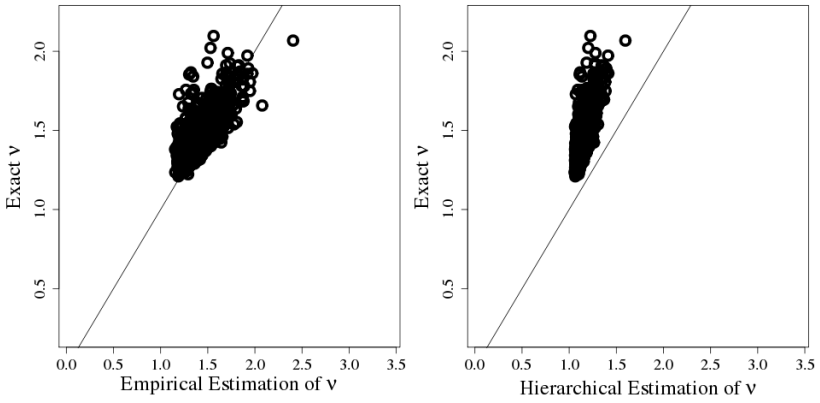

Fig. 5 Exact and asymptotic values of ν versus each other using hierarchical and empirical Bayes estimates of the model parameters

We are also interested to evaluate the performance of different methods in estimating the true area-specific covariate. In order to find the HB and CHB estimates of the x_i's, we apply the informative prior distributions on the model parameters. We use the MM estimates of the model parameters introduced in Ghosh and Sinha (2007) and the empirical Bayes estimates of μ and τ^2 to define the prior distribution on ψ, μ, and τ^2. To this end, we consider the prior distributions on the standard deviations, σ_e, σ_u, and σ_η as the uniform distribution between zero and twice the GS estimates of the model parameters. For τ, we consider a uniform distribution between zero and twice the empirical Bayes estimate of τ obtained from Efron and Morris (1975). For b_0, b_1, and μ, we consider normal prior distributions with the means equal to GS estimates of b_0 and b_1 and the empirical Bayes estimate of μ, respectively, and the corresponding variances as 4, 4, and 9, respectively. It is worth mentioning that the simulation studies (not shown here) confirm that the hierarchical model is quite sensitive to non-informative prior distributions and also to the

choice of hyper parameters. In order to address this problem, we consider the GS and EB estimates of the model parameters to define the priors. The densities of the estimated values of the true area-specific covariate using the Bayes, CB, EB, CEB, HB, and CHB methods are shown in Figure 6. As we expect, the CB, CEB, and the CHB methods lead to an ensemble of estimates with a histogram more similar to the true histogram of x_i's in comparison with the Bayes, EB, and HB methods, respectively. As Figure 6 shows, the CHB gives the closest ensemble of the estimates of the x_i's to the true area-specific covariate because the CHB makes use of the data twice - once in deriving the prior distribution using the GS estimates of the model parameters, and the second time in analyzing data.

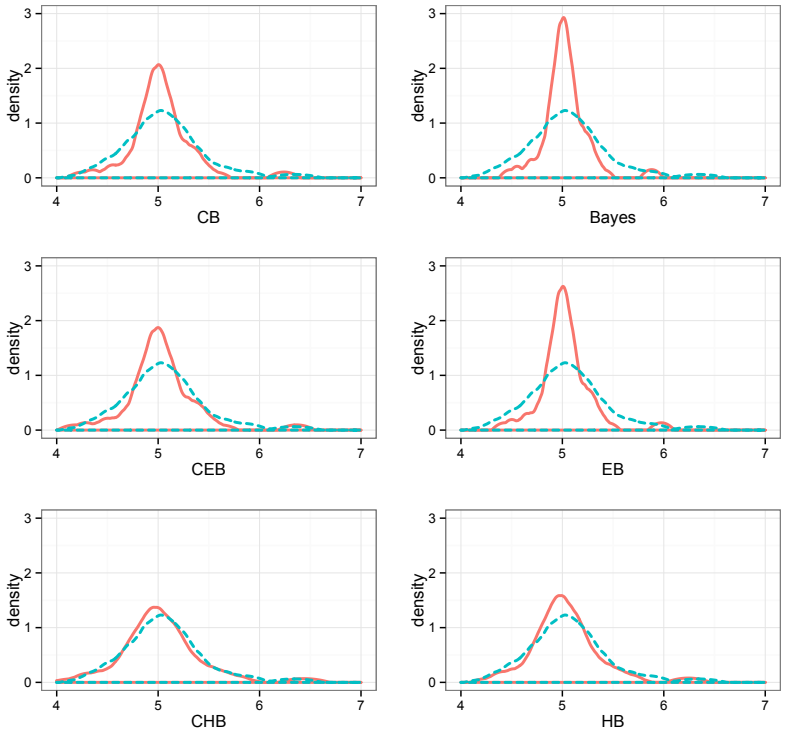

Fig. 6 The density of the estimated values of the true area-specific covariate using different approaches (solid line) vs the density of the true area-specific covariate (dashed line)

We also evaluated the performance of different estimated values of the small area means using the performance measures obtained in Section 3. To this end, the Se, Sp, PPV, and NPV of the small area mean predictors are calculated for each simulation using the threshold of 80. We report the results averaged over $R = 5000$ simulations (Figure 7). The result illustrates that in terms of the Se and PPV measures, the CEB method performs the best. However,

when it comes to the Sp, the HB method has the best performance while the CEB also shows a reasonable behaviour. The ML estimator shows the worst performance in terms of the Se and PPV.

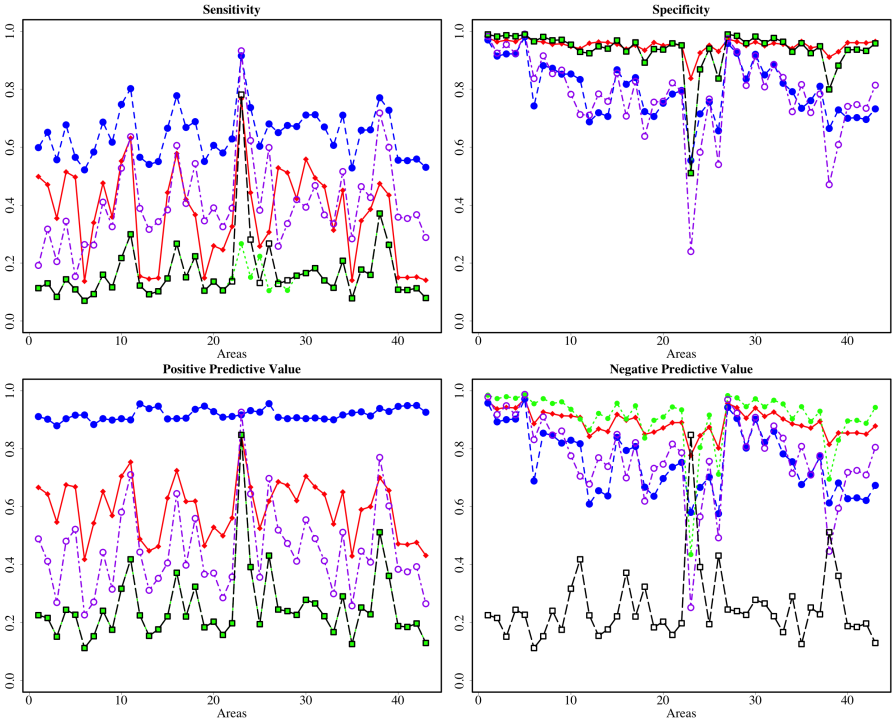

Fig. 7 The simulated values of the measures of performance of different areas: □—, ■—, ●--, ●· · ·, and ○-· are corresponding to $\hat{\gamma}_{ML}^{PEB}$, $\hat{\gamma}_{EB}^{PEB}$, $\hat{\gamma}_{CEB}^{PEB}$, $\hat{\gamma}_{HB}^{PEB}$, and $\hat{\gamma}_{CHB}^{PEB}$

The histograms of the predicted values of the first nine small area means over $R = 5000$ simulations are shown in Figure 8. The small area means are mostly below the threshold 80 such that the probability of getting small area means over 80 is small using the model parameters in the simulation study. Note that we observe similar behaviour in other small areas as well.

The MSPE of $\hat{\gamma}_{iCB}^{PB}$ is given by

$$\begin{aligned} \mathrm{E}(\hat{\gamma}_{iCB}^{PB} - \gamma_i)^2 &= \mathrm{E}(\hat{\gamma}_{iCB}^{PB} - \hat{\gamma}_i^B)^2 + \mathrm{E}(\hat{\gamma}_i^B - \gamma_i)^2 \\ &= (f_i B_i b_1)^2 \left[\mathrm{E}(x_{iCB} - x_i)^2\right] + f_i^2 \left(\sigma_e^2 \left(\frac{(1-B_i)^2}{n_i} + \frac{1}{N_i - n_i}\right) + B_i^2 \sigma_u^2\right), \end{aligned} \tag{6.1}$$

where (3.6) and (3.7) give $\mathrm{E}(x_{iCB})$ and $var(x_{iCB})$, respectively. Furthermore, Datta et al (2010) obtained the $MSPE(\hat{\gamma}_{iML}^{PB})$ as

$$MSPE(\hat{\gamma}_{iML}^{PB}) = \frac{f_i^2 \sigma_e^2 (1 - A_i)}{n_i} + \frac{1}{N_i} f_i \sigma_e^2,$$

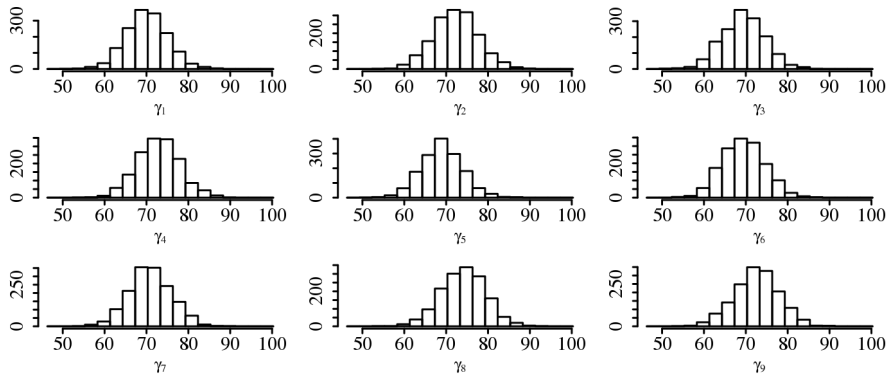

Fig. 8 The histograms of the predicted small area means for the first nine areas

where $A_i = \sigma_e^2/(\sigma_e^2 + n_i\sigma_u^2 + b_1^2\sigma_\eta^2)$. The $MSPE(\hat{\gamma}_{iB}^{PB})$ is also given by

$$\begin{aligned} &\mathrm{E}(\hat{\gamma}_{iB}^{PB} - \gamma_i)^2 = \mathrm{E}(\hat{\gamma}_{iB}^{PB} - \hat{\gamma}_i^B)^2 + \mathrm{E}(\hat{\gamma}_i^B - \gamma_i)^2 \\ &= (f_i B_i b_1)^2 \left[\mathrm{E}(x_{iB} - x_i)^2 + f_i^2 \left(\sigma_e^2 (\frac{(1-B_i)^2}{n_i} + \frac{1}{N_i - n_i}) + B_i^2 \sigma_u^2\right)\right., \quad (6.2) \end{aligned}$$

where $\mathrm{E}(x_{iB}) = C_i\mu + (1-C_i)x_i$ and $var(x_{iB}) = (1-C_i)^2\sigma_i^2$. In order to evaluate $MSPE(\hat{\gamma}_{iCB}^{PB})$, $MSPE(\hat{\gamma}_{iB}^{PB})$, and also $MSPE(\hat{\gamma}_{iML}^{PB})$, true values of the model parameters are used. Table 3 gives the MSPE of $\hat{\gamma}_{iB}^{PB}$, $\hat{\gamma}_{iCB}^{PB}$, $\hat{\gamma}_{iML}^{PB}$. As Table 3 shows, $\hat{\gamma}_{iB}^{PB}$ and $\hat{\gamma}_{iCB}^{PB}$ perform better than $\hat{\gamma}_{iML}^{PB}$ especially when the sample size is small. This happens because (2.5) and (2.7) use information from other areas to estimate the true area-specific covariate. In some areas, the MSPE of PEB estimators are smaller than the corresponding MSPE of the PB estimators. This is because there are cross-product terms involved in the MSPE of these estimators which are not negligible and indeed they are negative for some areas. We have planned to further investigate this issue in a separate research project when we deal with the MSPE estimation of the predictors of small area means.

Similar to Torkashvand et al (2015), we also assess the performance of predictors of small area means in terms of the empirical MSPE (EMSPE). The EMSPE of $\hat{\gamma}_i^{PEB}$ for different methods ($\hat{\gamma}_{iEB}^{PEB}$, $\hat{\gamma}_{iCEB}^{PEB}$, $\hat{\gamma}_{iHB}^{PEB}$, $\hat{\gamma}_{iCHB}^{PEB}$, and $\hat{\gamma}_{iML}^{PEB}$) is defined as

$$EMSPE(\hat{\gamma}_i^{PEB}) = \frac{1}{R}\sum_{r=1}^{R}(\hat{\gamma}_i^{PEB(r)} - \gamma_i^{(r)})^2.$$

In Torkashvand et al (2015), the superiority of $\hat{\gamma}_{iEB}^{PEB}$, the PEB predictor of the small area mean based on the James-Stein estimate of the true area-specific covariate, in terms of the EMSPE over the $\hat{\gamma}_{iML}^{PEB}$ and some competitive predictors of the small area mean was shown. Table 3 gives the EMSPE of $\hat{\gamma}_{iEB}^{PEB}$, $\hat{\gamma}_{iCEB}^{PEB}$, $\hat{\gamma}_{iHB}^{PEB}$, $\hat{\gamma}_{iCHB}^{PEB}$, and $\hat{\gamma}_{iML}^{PEB}$. Our findings indicate that the PEB

of small area means based on ML estimates of x_i's has the largest EMSPE while $\hat{\gamma}_{iHB}^{PEB}$ gives the minimum one in most areas. $\hat{\gamma}_{iCEB}^{PEB}$ and $\hat{\gamma}_{iCHB}^{PEB}$ show similar behaviour to $\hat{\gamma}_{iEB}^{PEB}$ and $\hat{\gamma}_{iHB}^{PEB}$, respectively, in terms of the EMSPE. As we expected the EMSPE of $\hat{\gamma}_{iCEB}^{PEB}$ and $\hat{\gamma}_{iCHB}^{PEB}$ are larger than $\hat{\gamma}_{iEB}^{PEB}$ and $\hat{\gamma}_{HB}^{PEB}$.

Table 3 Numerical values of the MSPE of $\hat{\gamma}_{iB}^{PB}$, $\hat{\gamma}_{iCB}^{PB}$ and $\hat{\gamma}_{iML}^{PB}$ and the EMSPE of $\hat{\gamma}_{iEB}^{PEB}$, $\hat{\gamma}_{iCEB}^{PEB}$, $\hat{\gamma}_{iHB}^{PEB}$, $\hat{\gamma}_{iCHB}^{PEB}$, and $\hat{\gamma}_{iML}^{PEB}$

		MSPE			*EMSPE*				
Area i	n_i	$\hat{\gamma}_{iB}^{PB}$	$\hat{\gamma}_{iCB}^{PB}$	$\hat{\gamma}_{iML}^{PB}$	$\hat{\gamma}_{iEB}^{PEB}$	$\hat{\gamma}_{iCEB}^{PEB}$	$\hat{\gamma}_{iHB}^{PEB}$	$\hat{\gamma}_{iCHB}^{PEB}$	$\hat{\gamma}_{iML}^{PEB}$
1	13	5.89	5.89	6.98	6.21	6.53	8.44	8.36	6.30
2	8	8.73	8.74	11.28	8.83	9.25	11.15	11.34	9.36
3	5	12.32	12.48	17.94	13.97	13.86	16.71	16.69	14.59
4	10	7.32	7.32	9.05	7.29	7.57	9.76	9.75	7.54
5	15	5.21	5.22	6.06	5.48	6.12	7.50	7.54	5.65
6	1	24.94	27.27	88.36	32.09	31.90	36.47	37.94	54.74
7	4	14.22	14.44	22.36	15.02	15.52	18.08	18.58	17.17
8	7	9.67	9.68	12.87	9.87	10.31	12.29	12.51	10.51
9	4	14.21	14.34	22.36	14.21	15.07	17.23	18.03	16.74
10	9	7.97	7.96	10.04	8.12	8.54	10.18	10.40	8.57
11	12	6.30	6.30	7.56	6.12	6.50	8.41	8.53	6.37
12	1	24.94	26.88	88.36	26.19	27.44	32.08	34.53	53.72
13	1	25.07	26.72	88.36	25.28	27.14	32.75	35.55	55.88
14	1	25.01	26.66	88.36	23.87	25.90	30.93	33.88	55.07
15	6	10.82	10.85	14.99	10.91	11.56	13.47	13.93	12.03
16	9	7.98	7.96	10.04	8.26	8.64	10.66	10.87	8.66
17	5	12.29	12.34	17.94	12.35	13.14	14.79	15.37	13.97
18	3	16.78	17.25	29.72	17.58	18.29	20.20	21.29	21.70
19	1	24.98	26.62	88.36	24.73	26.77	32.41	35.36	55.85
20	2	20.23	20.90	44.40	18.47	20.11	22.58	24.47	28.73
21	2	20.26	20.95	44.40	20.21	21.56	24.09	25.76	29.54
22	3	16.82	17.02	29.72	16.09	17.40	19.34	20.65	21.09
23	10	7.32	7.33	9.05	8.73	9.35	11.07	11.59	8.95
24	4	14.17	14.50	22.36	16.05	16.22	17.93	18.43	17.61
25	2	20.29	20.91	44.40	19.34	21.01	23.98	25.91	29.80
26	2	20.22	22.37	44.40	30.95	28.45	30.26	30.06	31.79
27	13	5.89	5.89	6.98	6.08	6.40	8.26	8.21	6.23
28	10	7.32	7.32	9.05	7.67	8.05	9.90	9.93	8.00
29	5	12.31	12.34	17.94	12.69	13.52	15.19	15.90	14.45
30	12	6.30	6.30	7.56	6.38	6.65	8.68	8.61	6.57
31	7	9.68	9.68	12.87	9.48	10.07	11.66	12.08	10.32
32	7	9.68	9.68	12.87	10.03	10.55	12.65	12.92	10.77
33	3	16.76	17.06	29.72	16.96	18.03	20.02	21.13	21.43
34	5	12.31	12.36	17.94	12.78	13.51	15.37	15.98	14.40
35	1	24.97	26.93	88.36	29.80	30.66	36.00	38.10	55.87
36	3	14.23	14.31	22.36	17.10	18.39	20.48	21.85	22.25
37	4	14.23	14.31	22.36	13.80	14.78	16.82	17.67	16.44
38	4	14.20	14.77	22.36	17.42	16.67	19.02	19.06	17.92
39	4	14.22	14.46	22.36	15.04	15.45	17.47	18.18	17.06
40	1	24.97	26.73	88.36	25.90	27.71	32.33	35.08	55.55
41	1	24.98	26.71	88.36	24.51	26.17	30.69	33.33	53.12
42	1	25.02	26.78	88.36	24.83	26.45	31.14	33.80	53.42
43	1	24.98	26.85	88.36	28.32	29.27	34.26	36.45	54.42

7. Concluding Remarks

Following Ghosh and Sinha (2007), Datta et al (2010), and Torkashvand et al (2015), a linear mixed model with the functional measurement error in the true area-specific covariate was considered in this paper. Following the general paradigm of Ghosh (1992), the constrained Bayes (CB) estimate of the true area-specific covariate was introduced in order to get a more precise estimate of the true underlying histogram of the population parameter. We showed that the CB estimate of the true area-specific covariate dominated the ML estimate in terms of the Bayes risk. Also, the pseudo-Bayes (PB) predictor of the small area means based on the CB estimate of the true area-specific covariate dominated the PB predictor of the small area means using the ML estimate. In order to evaluate the performance of different predictors of the small area means, the sensitivity (Se), specificity (Sp), positive predictive value (PPV), and negative predictive value (NPV) of the PB predictors of the small area means were obtained.

As an application, a cross-sectional data from the New Zealand population was analyzed. A simulation study was also conducted to evaluate the performance of the proposed approach. The simulation study showed that the histograms of the estimated values of the true area-specific covariates using the constrained Bayes (CB), the constrained empirical Bayes (CEB), and the constrained hierarchical Bayes (CHB) methods closely followed the true underlying histogram of the area-specific covariates. In addition, the pseudo empirical Bayes (PEB) predictor based on the CEB performed the best in terms of the Se and the PPV. We also observed desirable behaviour of our proposed estimators in terms of the Sp and the NPV measures. We noted that the PEB predictor based on the ML estimator of the true area-specific covariate performed the worst in terms of the Se and the PPV measures.

The MSPE and EMSPE of the different predictors of the small area means were also calculated. The MSPE of the small area mean predictor based on the constrained Bayes estimate of the true area-specific covariate dominates the MSPE of the small area mean predictor based on the ML estimator of the true area-specific covariate. Estimation of the MSPE using the jackknife method and comparison of different methods in terms of the relative bias, similar to Datta et al (2010) and Torkashvand et al (2015), remains as a future research project.

In this paper, we considered the case where we assume there is only one covariate in the study that is subject to the functional measurement error. There are many situations where there exists more than one covariate available in the study subject to the functional measurement error. Developing methodology for these situations can be considered as an extension of the current work. The other research project is to develop methodologies for the generalized linear mixed model with the covariate subject to the functional measurement error. This is of special interest when we are dealing with the logistic model and the goal is to estimate the probability of occurrence of a disease.

Acknowledgements We would like to thank three anonymous referees and the editor in chief for their constructive comments on an earlier version of the paper. Mohammad Jafari Jozani and Mahmoud Torabi gratefully acknowledge the research supports of the Natural Sciences and Engineering Research Council of Canada (NSERC). Elaheh Torkashvand's research is supported by the University of Manitoba Graduate Fellowship (UMGF) and Manitoba Graduate Scholarship (MGS).

References

Battese GE, Harter RM, Fuller WA (1988) An error-components model for prediction of county crop areas using survey and satellite data. J. Am. Stat. Assoc. 83(401):28–36

Carroll RJ, Ruppert, D, Stefanski LA, and Crainiceanu CM (2010) Measurement error in nonlinear models: a modern perspective. CRC press

Carroll RJ, Roeder K, Wasserman L (1999) Flexible parametric measurement error models. Biometrics 55(1):44–54

Datta GS, Ghosh M, Steorts R, Maples J (2011) Bayesian benchmarking with applications to small area estimation. Test 20 (3):574–588

Datta GS, Kubokawa T, Molina I, Rao JNK (2011) Estimation of mean squared error of model-based small area estimators. Test 20 (2): 367–388

Datta GS, Rao JNK, Torabi M (2010) Pseudo-empirical bayes estimation of small area means under a nested error linear regression model with functional measurement errors. J. Stat. Plan. Inference 140(11):2952–2962

Efron B, Morris C (1975) Data analysis using stein's estimator and its generalizations. J. Am. Stat. Assoc. 70 (350):311–319

Fuller WA (2009) Measurement error models. Vol. 305. John Wiley & Sons

Ghosh M (1992) Constrained bayes estimation with applications. J. Am. Stat. Assoc. 87(418):533–540

Ghosh M, Maiti T (1999) Adjusted bayes estimators with applications to small area estimation. Sankhyā Ser. B pp 71–90

Ghosh M, Sinha K (2007) Empirical bayes estimation in finite population sampling under functional measurement error models. J. Stat. Plan. Inference 137(9):2759–2773

Ghosh M, Steorts RC (2013) Two-stage benchmarking as applied to small area estimation. Test 22(4): 670–687

Ha NS (2013) Hierarchical Bayesian estimation of small area means using complex survey data. PhD dissertation, University of Maryland, College Park

Ha NS, Lahiri P (2014) Comments on: Single and two-stage cross-sectional and time series benchmarking procedures for small area estimation. SSBM 23(4): 670–673

Jiang J, Lahiri P (2006) Mixed model prediction and small area estimation. Test 15(1): 1–96

Jiang J (2010) Large sample techniques for statistics. SSBM

Kubokawa T, Strawderman WE (2013) Dominance properties of constrained Bayes and empirical Bayes estimators. Bernoulli 19(5B): 2200–2221

Lahiri P (1990) Adjusted Bayes and Empirical Bayes Estimation in Finite Population Sampling. Sankhyā Ser. B: 50–66

Louis TA (1984) Estimating a population of parameter values using bayes and empirical bayes methods. J. Am. Stat. Assoc. 79(386):393–398

Lyles RH, Xu J (1999) Classifying individuals based on predictors of random effects. Stat. Med. 18(1):35–52

Lyles RH, Manatunga AK, Moore RH, DuBois Bowman F, Cook CB (2007) Improving point predictions of random effects for subjects at high risk. Stat. Med. 26(6):1285–1300

Pfeffermann D, Sikov A, Tiller R (2014) Single- and two-stage cross-sectional and time series benchmarking procedures for small area estimation. Test 23(4): 631–666

Prasad NGN, Rao JNK (1990) The estimation of the mean squared error of small-area estimators. J. Am. Stat. Assoc. 85(409): 163–171

Rueda C, Menéndez JA, Gómez F (2010) Small area estimators based on restricted mixed models. Test 19(3): 558–579

Rao JNK, Molina I (2015) Small area estimation, Second Edition. Wiley, New York

Spjøtvoll E, Thomsen I (1987) Application of some empirical bayes methods to small area statistics. B. Int. Statist. Inst. 4:435–450

Torabi M (2011) Small area estimation using survey weights with functional measurement error in the covariate. Aust. NZ. J. Stat. 53(2):141–155

Torkashvand E, Jafari Jozani M, Torabi M (2015) Pseudo-empirical bayes estimation of small area means based on the james-stein estimation in linear regression models with functional measurement errors. Can. J. Stat. 43:265–287

Ugarte MD, Militino AF, Goicoa T (2009) Benchmarked estimates in small areas using linear mixed models with restrictions. Test 18(2): 342–364

Zhang W, Li N (2011) Prevalence, risk factors, and management of prehypertension. Int. J. Hypertens. 2011: 605359