Aust. N. Z. J. Stat. 2021 doi: 10.1111/j.1467-842X.XXX

Small area estimation under a semi-parametric covariate measured with error

Reyhane Sefidkar[1,2], Mahmoud Torabi[3]*and Amir Kavousi[2]

Shahid Sadoughi University of Medical Sciences, Shahid Beheshti University of Medical Sciences and University of Manitoba

Summary

In recent years, small area estimation has played an important role in statistics as it deals with the problem of obtaining reliable estimates for parameters of interest in areas with small or even zero sample sizes corresponding to population sizes. Nested error linear regression models are often used in small area estimation assuming that the covariates are measured without error and also the relationship between covariates and response variable is linear. Small area models have also been extended to the case in which a linear relationship may not hold, using penalised spline (P-spline) regression, but assuming that the covariates are measured without error. Recently, a nested error regression model using a P-spline regression model, for the fixed part of the model, has been studied assuming measurement error in covariate in the Bayesian framework. In this paper, we propose a frequentist approach to study a semi-parametric nested error regression model using P-spline with a covariate measured with error. In particular, the pseudo-empirical best predictors of small-area means and their corresponding mean squared prediction error estimates are studied. Performance of the proposed approach is evaluated through a simulation and also by a real data application.

Key words: jackknife; linear mixed model; mean squared prediction error; penalised spline

1. Introduction

Sample surveys have been long used as a preferred means of gathering information about a large population instead of census. Sometimes, to have a detailed analysis, estimating the parameters for sub-populations within the overall population of interest is needed, but, due

*Author to whom correspondence should be addressed.
[1] Center for Healthcare Data Modeling, Departments of Biostatistics and Epidemiology, School of Public Health, Shahid Sadoughi University of Medical Sciences, Yazd, Iran
[2] Department of Biostatistics, Shahid Beheshti University of Medical Sciences, Tehran, Iran
[3] Department of Community Health Sciences, University of Manitoba, Winnipeg, MB R3E 0W3, Canada
Email: Mahmoud.Torabi@umanitoba.ca

Acknowledgment. Constructive comments and suggestions of the Editor, an Associate Editor and two referees, which led to an improved version of this paper, are greatly appreciated. The research of Mahmoud Torabi was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) grant 368462-RGPIN.

to the cost and operational considerations, it is not always possible to have large enough sample size to warrant accurate estimates for those sub-populations which are called small areas. Since the traditional direct estimators do not provide adequate precision due to small sample sizes corresponding to population sizes, the demand to use and develop small area estimation methods has been greatly grown in recent years (Rao & Molina 2015). In order to provide reliable estimates for areas with small or even zero sample sizes, "indirect" estimators have been proposed in the context of small area estimation. The idea behind the indirect estimators is to increase the effective sample size by borrowing strength from other sources through a linking model using auxiliary information such as census data and administrative data (Pfeffermann et al. 2013; Rao & Molina 2015; Jiang 2017).

Small area models, which are based on mixed model methodology, are divided into two broad classes: (i) area-level models that relate the small area information on the response variable to area-specific auxiliary variables, and (ii) unit-level models which relate the unit values of the response variable to the unit-specific auxiliary variables with known area means and area-specific covariates. Rao & Molina (2015) gave an extensive review of model-based small area estimation under area-level and unit-level models. Focus of the current paper is on the unit-level models.

One of the basic assumptions in unit-level models is that the covariates are measured without error while this assumption may not be held in many real applications. Ignoring the measurement error (ME) may cause the small area predictors perform worse than direct estimators (Ybarra & Lohr 2008). In the context of classical ME model, there are two types of ME models, functional and structural ME model. In the functional type, the unknown true values of the covariate with ME are considered to be fixed which is in contrast with the structural type where the unobserved covariate is assumed to be stochastic. Ghosh & Sinha (2007), Datta, Rao & Torabi (2010), Torabi (2011) and Torkashvand, Jafari Jozani & Torabi (2015) studied the functional ME for an area-specific covariate in the nested error linear regression model. In these papers, the aim was to predict small area means with taking into account functional ME in covariate. To estimate the ME in covariate, Ghosh & Sinha (2007) proposed a moment estimator, Datta, Rao & Torabi (2010) suggested a maximum likelihood estimator (MLE) and Torkashvand, Jafari Jozani & Torabi (2015) used a James-Stein estimator to obtain pseudo-empirical Bayes (PEB) predictors of small area means. Another basic assumption in the unit level model is that the mean of the continuous outcome variable depends on the covariate value in a linear manner, while it might not hold in practice and due to complexity of the relationship, assuming a linear trend might only be a crude approximation. In such circumstances, parametric approaches will not properly work and to express this relation, a semi-parametric smoothing method such as penalized spline (P-spline) regression may be a good alternative (Eilers & Marx 1996). To see further applications of P-spline, we refer readers to the overview of P-spline models written by Ruppert, Wand & Carroll (2003).

In the context of small area estimation, Opsomer et al. (2008) studied P-spline regression model in the linear mixed model set-up. Torabi & Shokoohi (2015) extended Opsomer et al. (2008) model to a generalized linear mixed model (GLMM) to study normal and non-normal responses. Shokoohi & Torabi (2018) studied the P-spline regression model in the class of GLMMs to handle both time-series and cross-sectional response. Besides the P-spline model, the non-parametric M-quantile regression has been also studied to model the non-linear relationship between the qth M-quantile and the covariates in small area estimation (Pratesi, Ranalli & Salvati 2008, 2009; Salvati, Ranalli & Pratesi 2011). Jiang, Nguyen & Rao (2010) also proposed a procedure to select the small area model from a class of approximating splines, using a fence method.

In practical applications, however, there are many situations in which not only the predictor variable is not measured without error, but also the relationship between the response and the covariate is not linear or it is even hard to find the relationship between the response variable and the covariate. To deal with this problem, Hwang & Kim (2010, 2015) introduced a non-parametric nested error regression model with truncated polynomial basis functions and radial basis functions under functional ME model and predicted the small area means via a Bayesian approach. Hwang & Kim (2016) extended their non-parametric model by accommodating the covariates with and without ME again in a Bayesian framework.

In this paper, our aim is to take into account the functional ME in covariate in a semi-parametric nested error regression model from a frequentist perspective. To that end, in Section 2, we first rigorously study the model and present the "best" predictor of small area means, which is the best linear unbiased predictor. We then estimate the true covariate, using the maximum likelihood (ML) approach, to obtain the pseudo-"best" (PB) predictor of small area means. We also obtain mean squared prediction error (MSPE) of PB predictor of small area means. Furthermore, we use method-of-moments to estimate the model parameters to derive pseudo-empirical "best" (PEB) predictor of small area means. To estimate the MSPE of PEB predictor of small area means, we use the jackknife method. In order to evaluate our proposed PEB predictor and its corresponding jackknife MSPE estimator, a simulation study is conducted in Section 3. In Section 4, we employ the proposed model to predict the domain (area) mean blood pressure measured in National Health and Nutrition Examination Survey (NHANES), based on the cholesterol measured with error for some predefined domains, which is an important national source of information examining the health status of the population of the United States. Finally, we provide some concluding remarks in Section 5.

2. Model description

The nested error model with P-spline regression can be described as follows. Let y_{ij} be the variable of interest for the j-th unit ($j = 1, \ldots, N_i$) at the i-th small area $(i = 1, \ldots, m)$ with corresponding observed covariate w_{ij} as

$$y_{ij} = f_0(x_i) + \nu_i + e_{ij} \quad (i = 1, \ldots, m; j = 1, \ldots, N_i), \tag{1}$$

$$w_{ij} = x_i + \eta_{ij} \quad (i = 1, \ldots, m; j = 1, \ldots, N_i), \tag{2}$$

where N_i is the population size of i-th area, x_i is a continuous covariate which is fixed but *unknown*, $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_m)^\top$ is the area-level random effects with $\nu_i \overset{iid}{\sim} N(0, \sigma_\nu^2)$, e_{ij} is the random error with $e_{ij} \overset{iid}{\sim} N(0, \sigma_e^2)$, η_{ij} is ME with $\eta_{ij} \overset{iid}{\sim} N(0, \sigma_\eta^2)$, and the function $f_0(x_i)$ is generally unknown. Note that in the context of ME, we do not observe x_i, but rather we observe w_{ij} as in (2). We can approximate $f_0(x_i)$ sufficiently well using P-spline as

$$f_0(x_i) = b_0 + b_1 x_i + \ldots + b_p x_i^p + \sum_{a=1}^{k} \gamma_a (x_i - \tau_a)_+^p, \tag{3}$$

where p is the degree of spline, $(x)_+^p$ denotes the function $x^p I_{\{x>0\}}$, with I as the indicator function, $\{\tau_1, \ldots, \tau_k\}$ is a set of knots which ties a sequence of line segments to trace the continuous relation between the covariate and the response variable, $\boldsymbol{b} = (b_0, \ldots, b_p)^\top$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_k)^\top$ are the regression coefficients of parameters and P-spline parts of the model, respectively. It is assumed that $\gamma_a \overset{iid}{\sim} N(0, \sigma_\gamma^2)$ and ν_i, e_{ij}, η_{ij} and γ_a are assumed to be mutually independent. Considering k large enough and defining the knots in a way that they vastly spread out over the range of x_i, this class of approximation is very comprehensive and can approximate most smooth functions. In this study, we determine the number of spline knots (k) as the minimum of 40 and number of x_i's divided by 4, and the knots are quantiles of the distribution of x_i that are equally spaced (Ruppert 2002).

The goal is to predict the means of the response variable for the small areas of interest that is given by

$$\theta_i = \frac{1}{N_i} \sum_{j=1}^{N_i} y_{ij} \quad (i = 1, \ldots, m),$$

on the basis of the sample data which are denoted by $\{(y_{ij}, w_{ij}); j = 1, \ldots, n_i; i = 1, \ldots, m\}$, where n_i is the sample size of i-th small area. It is assumed that the models (1) and (2) hold for the sample data assuming that there is no sample selection bias and the sampling design is not informative.

Clearly, as each of the P-spline and the small area models are in the class of random-effects models, the combination of these two models can also be treated as a linear mixed effects model as

$$\boldsymbol{y} \;=\; \boldsymbol{X}\boldsymbol{b} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{D}\boldsymbol{\nu} + \boldsymbol{e},$$

where $\boldsymbol{y} = (\boldsymbol{y}_1{}^\top, \ldots, \boldsymbol{y}_m{}^\top)^\top$, $\boldsymbol{y_i} = (y_{i1}, \ldots, y_{in_i})^\top$, and $n_T = \sum_{i=1}^m n_i$ is the total sample size. We define $\boldsymbol{X} = (\boldsymbol{x}_{11} \ldots, \boldsymbol{x}_{1n_1}, \ldots, \boldsymbol{x}_{m1}, \ldots, \boldsymbol{x}_{mn_m})^\top$ and $\boldsymbol{Z} = (\boldsymbol{z}_{11}, \ldots, \boldsymbol{z}_{1n_1}, \ldots, \boldsymbol{z}_{m1}, \ldots, \boldsymbol{z}_{mn_m})^\top$ where $\boldsymbol{x}_{ij} = \boldsymbol{x}_i = (1, x_i, \ldots, x_i^p)^\top$ and $\boldsymbol{z}_{ij} = \boldsymbol{z}_i = ((x_i - \tau_1)_+^p, \ldots, (x_i - \tau_k)_+^p)^\top$ are the vectors of the covariates for each sample in i-th area, respectively, for $j = 1, \ldots, n_i$. We also define $\boldsymbol{D} = (\boldsymbol{d}_{11}, \ldots, \boldsymbol{d}_{1n_1}, \ldots, \boldsymbol{d}_{m1}, \ldots, \boldsymbol{d}_{mn_m})^\top$ where $\boldsymbol{d}_{ij} = \boldsymbol{d}_i = (0, \ldots, 0, 1, 0, \ldots, 0)^\top$ which the i-th element is equal to 1 and $\boldsymbol{e} = (\boldsymbol{e}_1^\top, \ldots, \boldsymbol{e}_m^\top)^\top$, $\boldsymbol{e_i} = (e_{i1}, \ldots, e_{in_i})^\top$.

2.1. Best predictor

To predict the i-th small area mean, first, assume that the value of the covariate x_i is not subject to the ME. Then, using the observed response data, the best predictor is given by

$$\begin{aligned}
\hat{\theta}_i^B(x_i, \boldsymbol{\phi}_1) \;&=\; N_i^{-1}\Big[\sum_{j=1}^{n_i} y_{ij} + \sum_{j=n_i+1}^{N_i} \hat{y}_{ij}\Big] \\
&=\; (1 - f_i)\bar{y}_i + f_i\Big(b_0 + b_1 x_i + \ldots + b_p x_i^p + \sum_{a=1}^{k} \hat{\gamma}_a (x_i - \tau_a)_+^p + \hat{\nu}_i\Big),
\end{aligned}$$

where $f_i = 1 - n_i/N_i$ for $i = 1, \ldots, m$, $\hat{\boldsymbol{\gamma}} = \boldsymbol{\Sigma}_\gamma \boldsymbol{Z}^\top \boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})$, $\hat{\boldsymbol{\nu}} = \boldsymbol{\Sigma}_\nu \boldsymbol{D}^\top \boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})$, which are the BLUP of the random effects $\boldsymbol{\gamma}$ and $\boldsymbol{\nu}$, respectively, where $\mathbf{V} = \mathbf{var}(\mathbf{y}) = \mathbf{Z}\boldsymbol{\Sigma}_\gamma \mathbf{Z}^\top + \mathbf{D}\boldsymbol{\Sigma}_\nu \mathbf{D}^\top + \boldsymbol{\Sigma}_\mathbf{e}$ with $\boldsymbol{\Sigma}_\gamma = \sigma_\gamma^2 \boldsymbol{I}_k$, $\boldsymbol{\Sigma}_\nu = \sigma_\nu^2 \boldsymbol{I}_m$, and $\boldsymbol{\Sigma}_e = \sigma_e^2 \boldsymbol{I}_{n_T}$ and $\boldsymbol{\phi}_1 = (\mathbf{b}^\top, \sigma_\mathbf{e}^\mathbf{2}, \sigma_\nu^\mathbf{2}, \sigma_\gamma^\mathbf{2})$ is assumed to be known. The corresponding $MSPE$ of $\hat{\theta}_i^B$ is then given by

$$E(\hat{\theta}_i^B - \theta_i)^2 = f_i^2 \boldsymbol{q}_i^\top (\boldsymbol{\Sigma}_s - \boldsymbol{\Sigma}_s \boldsymbol{\Omega} \mathbf{V}^{-1} \boldsymbol{\Omega}^\top \boldsymbol{\Sigma}_s) \boldsymbol{q}_i,$$

where $\boldsymbol{\Omega} = (\mathbf{Z}, \mathbf{D})^\top$, $\mathbf{s} = (\boldsymbol{\gamma}^\top, \boldsymbol{\nu}^\top)^\top$, $\mathbf{q}_i = (\mathbf{z}_i^\top, \mathbf{l}_i^\top)^\top$, where $\mathbf{l}_i$ is a vector with one as the i-th element and zero in other $(m-1)$ elements, and $\boldsymbol{\Sigma}_s = \begin{bmatrix} \boldsymbol{\Sigma}_\gamma & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_\nu \end{bmatrix}$.

As it is clear, the introduced best predictor depends on x_i which may be measured with error in practice. As a result, the true values are not observed in such occasions. In the case of nested error linear regression model, Ghosh & Sinha (2007) estimated x_i with its moment estimator, $\bar{w}_i$. For the same set-up, Datta, Rao & Torabi (2010) proposed a MLE for x_i using all the available data $\{y_{ij}, w_{ij}; j = 1, \ldots, n_i; i = 1, \ldots, m\}$ which is more efficient than Ghosh & Sinha (2007) estimator in terms of $MSPE$ of small area predictors. Following Datta, Rao & Torabi (2010), we use the MLE method to estimate the true value of the covariate x_i.

2.2. Pseudo-best predictor

In order to estimate x_i, we consider the sample means of our data $\{(y_{ij}, w_{ij}); j = 1, \ldots, n_i; i = 1, \ldots, m\}$ through the equations (1)-(3) as

$$\begin{aligned} \bar{y}_i &= b_0 + b_1 x_i + \ldots + b_p x_i^p + \sum_{a=1}^{k} \gamma_a (x_i - \tau_a)_+^p + \nu_i + \bar{e}_i \\ &= b_0 + b_1 x_i + \ldots + b_p x_i^p + \bar{u}_{1i}, \qquad (4) \\ \bar{w}_i &= x_i + \bar{u}_{2i}, \qquad (5) \end{aligned}$$

where $\bar{u}_{1i} \sim N\Big(0, \sigma_\gamma^2 \sum_{a=1}^{k} (x_i - \tau_a)_+^{2p} + \sigma_\nu^2 + \frac{\sigma_e^2}{n_i}\Big)$ and $\bar{u}_{2i} = \bar{\eta}_i \sim N(0, \frac{\sigma_\eta^2}{n_i})$. Since $\bar{u}_{1i}$ is independent of $\bar{u}_{2i}$, the log-likelihood function, $l(x_i)$, can be expressed as the log of joint density $f(\bar{y}_i, \bar{w}_i | x_i) = f(\bar{y}_i | x_i) f(\bar{w}_i | x_i)$ through

$$\begin{aligned} l(x_i) &= \log(f(\bar{y}_i, \bar{w}_i | x_i)) \\ &\propto -\frac{1}{2\sigma_\eta^2} n_i (\bar{w}_i - x_i)^2 - \frac{1}{2}\, log\Big[2\pi\Big(\sigma_\gamma^2 \sum_{a=1}^{k} (x_i - \tau_a)_+^{2p} + \sigma_\nu^2 + \frac{\sigma_e^2}{n_i}\Big)\Big] \\ &\quad - \frac{(\bar{y}_i - b_0 - b_1 x_i - \ldots - b_p x_i^p)^2}{2\Big[\sigma_\gamma^2 \sum_{a=1}^{k} (x_i - \tau_a)_+^{2p} + \sigma_\nu^2 + \frac{\sigma_e^2}{n_i}\Big]}. \end{aligned}$$

Since $l(x_i)$ does not have a closed form, we use numerical methods for maximization. Maximizing the likelihood function with respect to x_i and substituting $\tilde{x}_i$, which is the estimate of x_i, for x_i in the best estimator leads to the following PB predictor

$$\begin{aligned} \hat{\theta}_i^{PB} &= \hat{\theta}_i^{PB}(\boldsymbol{\phi}) \\ &= (1 - f_i)\bar{y}_i + f_i (b_0 + b_1 \tilde{x}_i + \ldots + b_p \tilde{x}_i^p + \sum_{a=1}^{k} \hat{\gamma}_a (\tilde{x}_i - \tau_a)_+^p + \hat{\nu}_i), \end{aligned}$$

where $\boldsymbol{\phi} = (\boldsymbol{\phi}_1, \sigma_\eta^2)$. Since $E(\hat{\theta}_i^B - \theta_i | \mathbf{y}_i) = 0$, the $MSPE$ of PB predictor is given by

$$\begin{aligned} MSPE(\hat{\theta}_i^{PB}) &= E(\hat{\theta}_i^{PB} - \theta_i)^2 \\ &= f_i^2 E\Big\{ b_1(\tilde{x}_i - x_i) + \ldots + b_p(\tilde{x}_i^p - x_i^p) + \sum_{a=1}^{k} \hat{\gamma}_a \Big[(\tilde{x}_i - \tau_a)_+^p - (x_i - \tau_a)_+^p \Big] \Big\}^2 \\ &\quad + f_i^2 \boldsymbol{q}_i^\top (\boldsymbol{\Sigma}_s - \boldsymbol{\Sigma}_s \boldsymbol{\Omega} \mathbf{V}^{-1} \boldsymbol{\Omega}^\top \boldsymbol{\Sigma}_s) \boldsymbol{q}_i \equiv g_{1i}(\boldsymbol{\phi}). \end{aligned}$$

In reality, the PB predictor and corresponding $MSPE$ are not computable as they depend on the model parameters $\boldsymbol{\phi}$.

2.3. Pseudo-empirical best predictor

In order to predict the small area means we now need to estimate $\boldsymbol{\phi}$. Then substituting $\hat{\boldsymbol{\phi}}$ for $\boldsymbol{\phi}$ in $\hat{\theta}_i^{PB}$ gives PEB predictor, $\hat{\theta}_i^{PEB}$, of small area means. Here, we use the method-of-moments to estimate the model parameters $\boldsymbol{\phi}$. Following Ghosh & Sinha (2007), the estimates of random error and ME variances are

$$\hat{\sigma}_e^2 = \frac{SSW_{\mathbf{y}}}{n_T - m},$$
$$\hat{\sigma}_\eta^2 = \frac{SSW_{\mathbf{w}}}{n_T - m},$$

where

$$\begin{aligned} SSW_{\mathbf{y}} &= \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2, \\ SSW_{\mathbf{w}} &= \sum_{i=1}^{m}\sum_{j=1}^{n_i}(w_{ij} - \bar{w}_i)^2. \end{aligned}$$

In the next step, we need to estimate the regression coefficients $\mathbf{b}$. Based on equations (4)-(5), we can write the model as

$$\bar{y}_i = \bar{\boldsymbol{x}}_i^\top \boldsymbol{b} + \bar{\boldsymbol{z}}_i{}^\top \boldsymbol{\gamma} + \bar{\boldsymbol{d}}_i{}^\top \boldsymbol{\nu} + \bar{e}_i, \tag{6}$$
$$\boldsymbol{h}_i = \bar{\boldsymbol{x}}_i + \bar{\boldsymbol{\eta}}_i, \tag{7}$$

where $\mathbf{h}_i = (1, \bar{w}_i, \ldots, \bar{w}_i^p)^\top$ and $\bar{\boldsymbol{\eta}}_i = (\bar{\eta}_{0i}, \bar{\eta}_{1i}, \ldots, \bar{\eta}_{pi})^\top$. Since $E(\bar{e}_i) = 0$, $E(\bar{\boldsymbol{\eta}}_i) = \mathbf{0}$, $var(\bar{e}_i) = \sigma_e^2/n_i$, and $var(\bar{\boldsymbol{\eta}}_i) = n_i^{-1}\boldsymbol{\Sigma}_\eta$ with $\boldsymbol{\Sigma}_\eta$ as the variance-covariance matrix of measurement errors, we have

$$E(\boldsymbol{h}_i) = \bar{\boldsymbol{x}}_i,$$

159

$$E(n_i \boldsymbol{h}_i \boldsymbol{h}_i{}^\top) = n_i \bar{\boldsymbol{x}}_i \bar{\boldsymbol{x}}_i^\top + \boldsymbol{\Sigma}_\eta$$

From equation (6), we also have

$$n_i \bar{\boldsymbol{x}}_i \bar{y}_i = n_i \bar{\boldsymbol{x}}_i \bar{\boldsymbol{x}}_i^\top \boldsymbol{b} + n_i \bar{\boldsymbol{x}}_i \bar{\boldsymbol{z}}_i^\top \boldsymbol{\gamma} + n_i \bar{\boldsymbol{x}}_i \bar{\boldsymbol{d}}_i^\top \boldsymbol{\nu} + n_i \bar{\boldsymbol{x}}_i \bar{e}_i \tag{8}$$

and taking expectation and then taking expectation from both sides of the equation (8) leads to

$$\begin{aligned} n_i\bar{\boldsymbol{x}}_i E(\bar{y}_i) &= n_i\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^\top \boldsymbol{b}, \\ \frac{1}{m}\sum_{i=1}^m n_i\bar{\boldsymbol{x}}_i E(\bar{y}_i) &= \Big(\frac{1}{m}\sum_{i=1}^m n_i\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^\top\Big)\boldsymbol{b}, \\ \boldsymbol{b} &= \left(\frac{1}{m}\sum_{i=1}^m n_i\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^\top\right)^{-1}\left(\frac{1}{m}\sum_{i=1}^m n_i\bar{\boldsymbol{x}}_i E(\bar{y}_i)\right). \end{aligned}$$

The plug-in estimator of $\boldsymbol{b}$ is then given by

$$\hat{\boldsymbol{b}} = \left(\frac{1}{m}\sum_{i=1}^m (n_i\boldsymbol{h}_i\boldsymbol{h}_i{}^\top - \hat{\Sigma}_\eta)\right)^{-1}\left(\frac{1}{m}\sum_{i=1}^m n_i\boldsymbol{h}_i\bar{y}_i\right).$$

To estimate σ_γ^2 and σ_ν^2, let $MSB_{\mathbf{y}} = (m-1)^{-1}\sum_{i=1}^m n_i(\bar{y}_i - \bar{y})^2$, then, following Datta, Rao & Torabi (2010), we have

$$\begin{aligned} E(MSB_{\mathbf{y}}) &= \frac{\sigma_\gamma^2}{m-1}\Big[\sum_{i=1}^m (n_i - \frac{n_i^2}{n_T})\sum_{a=1}^k (x_i - \tau_a)_+^{2p}\Big] + \frac{\sigma_\nu^2}{m-1}\sum_{i=1}^m (n_i - \frac{n_i^2}{n_T}) + \sigma_e^2 + \\ &\quad \frac{1}{m-1}\sum_{i=1}^m n_i\Big[b_1(x_i - \bar{x}) + \ldots + b_p(x_i^p - \bar{x^p})\Big]^2, \end{aligned}$$

and the estimates of σ_γ^2 and σ_ν^2 are then obtained as

$$\begin{aligned} \hat{\sigma}_\gamma^2 &= \Bigg\{(m-1)MSB_{\mathbf{y}} - (m-1)\hat{\sigma}_e^2 - \sum_{i=1}^m n_i\Big[\hat{b}_1(\hat{x}_i - \bar{x}) + \ldots + \hat{b}_p(\hat{x}_i^p - \bar{x^p})\Big]^2 \\ &\quad -\hat{\sigma}_\nu^2\sum_{i=1}^m (n_i - \frac{n_i^2}{n_T})\Bigg\}\Big/\Bigg\{\sum_{i=1}^m (n_i - \frac{n_i^2}{n_T})\sum_{a=1}^k (\hat{x}_i - \tau_a)_+^{2p}\Bigg\}, \end{aligned}$$

and

$$\begin{aligned} \hat{\sigma}_\nu^2 &= \frac{1}{\sum_{i=1}^m (n_i - \frac{n_i^2}{n_T})}\Bigg\{(m-1)MSB_{\mathbf{y}} - (m-1)\hat{\sigma}_e^2 - \sum_{i=1}^m n_i\Big[\hat{b}_1(\hat{x}_i - \bar{x}) + \ldots + \hat{b}_p(\hat{x}_i^p - \bar{x^p})\Big]^2 \\ &\quad -\hat{\sigma}_\gamma^2\Big[\sum_{i=1}^m (n_i - \frac{n_i^2}{n_T})\sum_{a=1}^k (\hat{x}_i - \tau_a)_+^{2p}\Big]\Bigg\}, \end{aligned}$$

where $\bar{x^j} = \frac{1}{n_T}\sum_{i=1}^{m} n_i \hat{x}_i^j, (j = 1, \dots, p)$. Since the estimates of $(x_1, \dots, x_p, \sigma_\nu^2, \sigma_\gamma^2)$ are dependent of each other, they are estimated via an iterative algorithm. In particular, to estimate $(x_1, \dots, x_p)$ the above two equations related to $\hat{\sigma}_\gamma^2$ and $\hat{\sigma}_\nu^2$ are estimated iteratively with respect to $(x_1, \dots, x_p)$ through the within mean square error of the observed values of the covariate. Finally, the PEB predictor of θ_i is given by

$$\begin{aligned}\hat{\theta}_i^{PEB} &= \hat{\theta}_i^{PB}(\hat{\boldsymbol{\phi}}) \\ &= (1 - f_i)\bar{y}_i + f_i\Big(b_0 + b_1\tilde{x}_i + \ldots + b_p\tilde{x}_i^p + \sum_{a=1}^{k} \hat{\gamma}_a(\tilde{x}_i - \tau_a)_+^p + \hat{\nu}_i\Big).\end{aligned}$$

To measure the variability of $\hat{\theta}_i^{PEB}$, the $MSPE$ of PEB predictor can be decomposed as

$$\begin{aligned}MSPE(\hat{\theta}_i^{PEB}) &= E(\hat{\theta}_i^{PB} - \theta_i)^2 + E(\hat{\theta}_i^{PEB} - \hat{\theta}_i^{PB})^2 + 2E\Big[(\hat{\theta}_i^{PB} - \theta_i)(\hat{\theta}_i^{PEB} - \hat{\theta}_i^{PB})\Big] \\ &= M_{1i} + M_{2i} + 2M_{3i},\end{aligned}$$

where $M_{1i} = g_{1i}(\boldsymbol{\phi})$. The $MSPE$ of PEB predictor is not computable as it depends on model parameters. In order to estimate the $MSPE$ of PEB predictor of small area means, we apply the jackknife method which was proposed by Jiang et al. (2002) and Chen & Lahiri (2002). To be able to use the jackknife method, similar to other studies which were done in the context of functional ME for area-level and unit-level models such as Ybarra & Lohr (2008), Datta, Rao & Torabi (2010) and Torkashvand, Jafari Jozani & Torabi (2015), we approximate the $MSPE$ as

$$MSPE(\hat{\theta}_i^{PEB}) \approx M_{1i} + M_{2i},$$

by ignoring the cross-product term, noting that there is no-closed form expression for the $MSPE(\hat{\theta}_i^{PEB})$ (Haslett & Welsh 2019). We will report the magnitude of M_{3i} in the simulation study section.

To estimate M_{1i}, a jackknife bias correction is used which is given by

$$\hat{M}_{1iJ} = g_{1i}(\hat{\boldsymbol{\phi}}) - \sum_{l \neq i} \psi_l \Big[g_{1i}(\hat{\boldsymbol{\phi}}_{-l}) - g_{1i}(\hat{\boldsymbol{\phi}})\Big], \quad l = 1, \dots, m, \tag{9}$$

where $\psi_l = 1 + O(m^{-1})$ is a suitable weight (Chen & Lahiri 2002). Here, $g_{1i}(\hat{\boldsymbol{\phi}})$ is the plug-in estimator of $g_{1i}(\boldsymbol{\phi})$ and $\hat{\boldsymbol{\phi}}_{-l}$ is the moment estimator of $\boldsymbol{\phi}$, obtained by omitting the l-th area data set from the full data set $\{(y_{ij}, w_{ij}); j = 1, \dots, n_i; i = 1, \dots, m\}$. This is done for each $l \neq i$ (except the i-th area) to get $m - 1$ estimators for $\boldsymbol{\phi}$.

The jackknife estimator of M_{2i} is given by

$$\hat{M}_{2iJ} = \sum_{l \neq i} \psi_l (\hat{\theta}_{i,-l}^{PEB} - \hat{\theta}_i^{PEB})^2, \;\; l = 1, \ldots, m, \tag{10}$$

where $\hat{\theta}_{i,-l}^{PEB}$ is the plug-in estimator of $\hat{\theta}_i^{PB}$, in which the vector of the parameters ($\boldsymbol{\phi}$) is estimated by deleting the l-th area data set from the full data set each time. Finally, the jackknife estimator of $MSPE(\hat{\theta}_i^{PEB})$ is obtained by taking the sum of (9) and (10) as

$$mspe_J(\hat{\theta}_i^{PEB}) = \hat{M}_{1iJ} + \hat{M}_{2iJ}.$$

Assuming $\psi_l = 1 - \boldsymbol{h}_l^\top (\sum_{t \neq i} \boldsymbol{h}_t \boldsymbol{h}_t^\top)^{-1} \boldsymbol{h}_l$ and $\psi_l = \frac{m-2}{m-1}$, the weighted and unweighted versions of jackknife estimator of $MSPE(\hat{\theta}_i^{PEB})$ are obtained, respectively. Note that in small area estimation (Rao & Molina 2015), the notation $mspe$ is usually used as the estimator of $MSPE$.

3. Simulation study

In this section, we carry out a simulation study to compare the performance of the proposed approach in the P-spline model which takes into account the ME in the area level predictor variable and the P-spline model which ignores the ME (naive model). To this end, the population responses are generated from the model (1) with three choices for $f_0(x_i)$: linear, quadratic, and exponential 1 (see later of this section for details of choices of $f_0(x_i)$). Note that in small area estimation, the asymptotic result will apply for large number of small areas m. So, for large m the effects of model parameter estimate would be vanished as long as the model parameter estimators are consistent. Hence, it is important in the context of small area estimation to show how good is the proposed model for finite sample (small m). Therefore, we assume that the population units are distributed across $m = 40$ areas equally in a way that $N_i = 400$, $(i = 1, \ldots, m)$, and equal sample sizes are taken from each area as $n_i = 4, (i = 1, \ldots, m)$. We generate $R = 1000$ independent sets of $\{\nu_i^{(r)}; i = 1, \ldots, m\}$, $\{e_{ij}^{(r)}; j = 1, \ldots, N_i; i = 1, \ldots, m\}$ from Normal distribution with mean zero and variance σ_ν^2 and σ_e^2, respectively. We assume $\sigma_\nu^2 = 1$ and $\sigma_e^2 = 1$ for linear case and $\sigma_\nu^2 = 0.1$ and $\sigma_e^2 = 0.3$ for non-linear cases. The true values of the predictor $\{x_i; i = 1, \ldots, m\}$ are also generated from a uniform distribution between 10 and 30 for linear case and uniform distribution between -3 and 3 for non-linear cases and treat them fixed through the simulation study. Using $\{x_i, \nu_i^{(r)}, e_{ij}^{(r)}\}$, the population responses are generated from the model (1) with three choices of $f_0(x_i)$ (linear, quadratic, and exponential 1) as

$$\begin{aligned} y_{ij}^{(r)} &= 1 + x_i + \nu_i^{(r)} + e_{ij}^{(r)}, \;\; r = 1, \ldots, R; j = 1, \ldots, N_i; i = 1, \ldots, m, \\ y_{ij}^{(r)} &= 0.4 + 0.4x_i - 0.65x_i^2 + \nu_i^{(r)} + e_{ij}^{(r)}, \\ y_{ij}^{(r)} &= 1 + x_i - 0.7\, exp(x_i) + \nu_i^{(r)} + e_{ij}^{(r)}, \end{aligned}$$

following Breidt, Claeskens & Opsomer (2005), Rao, Sinha & Dumitrescu (2014) and Shokoohi & Torabi (2018). The population mean response of i-th area for r-th simulation is given by

$$\theta_i^{(r)} = \frac{1}{N_i} \sum_{j=1}^{N_i} y_{ij}^{(r)}.$$

We then generate simple random samples from each simulated population responses. Furthermore, we generate the observed values of the predictor variable from the ME model $w_{ij}^{(r)} = x_i + \eta_{ij}^{(r)}$, $(j = 1, \ldots, n_i; i = 1, \ldots, m)$, where $\eta_{ij}^{(r)}$ is generated from a normal distribution with mean zero and variance $\sigma_\eta^2 = 2$ for linear case and $\sigma_\eta^2 = 0.6$ for non-linear quadratic and exponential 1 cases. Thereafter, for each simulated data set $\{(w_{ij}^{(r)}, y_{ij}^{(r)}); j = 1, \ldots, n_i; i = 1, \ldots, m\}$ in each scenario (linear, quadratic, and exponential 1), the P-spline model with and without ME are fitted assuming $p = 1$ (which has piecewise linear fit, with the changes in slope at each knot regarded as random with variance σ_γ^2). Table 1 presents the moment estimators of the proposed model parameters for each scenario. In particular, in the case of linear model as the true model, the model provides the regression coefficients which are very close to the true values; note that the variance component of P-spline (σ_γ^2) is close to zero since the true model is linear. Also, in the cases of quadratic and exponential 1 models, the fitted model works very well to track the true values, note that in these two scenarios, we only need to compare the variance components estimates of the model with the corresponding true values. We observe that the variations of the models from the linearity (quadratic and exponential 1) are well captured through the estimate of σ_γ^2 in the proposed model.

[Table 1 about here]

Figures 1a, 1b, and 1c show PEB predictions of small area means. It seems that both models have similar predictions for small area means.

[Figure 1 about here]

Now we need to evaluate the accuracy of $\hat{\theta}_i^{PEB}$ in our proposed approach. To that end, we calculate the empirical $MSPE$ ($EMSPE$) of $\hat{\theta}_i^{PEB}$ which is given by

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

To also evaluate the magnitude of the cross-product term involved in the $MSPE$ of PEB predictor of small area means, $MSPE$ of $\hat{\theta}_i^{PEB(r)}$ can be decomposed as

$$EMSPE(\hat{\theta}_i^{PEB(r)}) = M_{ip} = M_{1ip} + M_{2ip} + 2M_{3ip},$$

where $M_{1ip} = R^{-1}\sum_{r=1}^{R}(\hat{\theta}_i^{PB(r)} - \theta_i^{(r)})^2$, $M_{2ip} = R^{-1}\sum_{r=1}^{R}(\hat{\theta}_i^{PEB(r)} - \hat{\theta}_i^{PB(r)})^2$ and $M_{3ip} = R^{-1}\sum_{r=1}^{R}(\hat{\theta}_i^{PB(r)} - \theta_i^{(r)})(\hat{\theta}_i^{PEB(r)} - \hat{\theta}_i^{PB(r)})$.

[Figure 2 about here]

[Figure 3 about here]

[Figure 4 about here]

Figures 2a, 3a, and 4a show the $EMSPE$ of $\hat{\theta}_i^{PEB(r)}$ and its decomposition for the both proposed and naive models for three cases (linear, quadratic, and exponential 1). Based on the results for the proposed and naive models for all three cases (linear, quadratic, and exponential 1), it seems that the PEB predictions of small area means are near the true means since the values of M_{ip} are close to zero particularly in the cases of quadratic and exponential forms, and the most contribution of $MSPE$ is attributed to M_{1ip} as expected. It is also clear from Figures 2a, 3a, and 4a that the contribution of the cross-product M_{3ip} term involved in the $MSPE$ is small in the proposed model compared to the naive model.

In order to evaluate the performance of the weighted, $mspe_{JW}(\hat{\theta}_i^{PEB})$, and unweighted, $mspe_J(\hat{\theta}_i^{PEB})$, jackknife $MSPE$ estimators, the empirical relative bias (RB) of these estimators are computed through the following expression

$$RB_i = \frac{E(mspe_i)}{EMSPE_i} - 1 \quad (i = 1, \ldots, m),$$

where $E(mspe_i)$ is the average of simulated jackknife $MSPE$ estimate of PEB predictor of small area mean i. The results of RB of jackknife $mspe$ of PEB predictor of small area means (weighted and unweighted) for the proposed and naive models are also reported for all three cases (linear, quadratic, and exponential 1) in Figures 2b, 3b, and 4b. As we expect, the proposed P-spline model performs very well in terms of RB for the three scenarios in this simulation set-up, and it appears that the unweighted $MSPE$ estimates perform as well as the corresponding weighted version. Based on our empirical findings, jackknife method causes serious overestimation of $MSPE$ of the PEB predictor in naive model due to the large values of M_{3ip} as shown in Figures 2a to 4a. Furthermore, in order to evaluate the effect of level of measurement error on estimates and predictions, we consider different values of measurement error variance for linear and non-linear cases. Table 2 presents the moment estimators of the proposed model parameters for each scenario. As shown in Table 2, our proposed P-spline model works very well in terms of model parameters estimate for different spline forms and measurement error variances.

[Table 2 about here]

In terms of performance of naive model, for example, we observed that in the case of linear model, with increasing the value of measurement error variance, parameter estimate of slope is attenuated (not shown here) unlike the proposed P-spline model. The bias in the slope estimate which is caused by measurement error is discussed in the literature (Carroll 2006). Figures 4 to 6 show $EMSPE$ of PEB predictors and its components for the proposed and naive models and related percent relative bias of jackknife estimators of unweighted and weighted $MSPE$ for the proposed and naive models. With increasing measurement error variance σ_η^2, the proposed P-spline model still shows good performance in terms of RB in all three scenarios. Note that in case of naive model with exponential form, the cross-product term (M_{3in}) has the same magnitude as the leading term (M_{1in}) but with opposite sign. This is the reason that RB for the naive model is also as good as the proposed model in the case of exponential model.

[Figure 5 about here]

[Figure 6 about here]

[Figure 7 about here]

4. Application

In this section, we employ our proposed P-spline model to analyze data from the 2013–2014 US NHANES. The NHANES is a yearly survey to determine the health and nutritional status of adults and children in the United States. According to the literature, there is a significant positive relationship between obesity and blood pressure (Lee, Bacha & Arslanian 2006; Choy et al. 2011; Duncan et al. 2013). Since waist circumference (WC) index is expressed as the main indicator of abdominal fat accumulation, hence, in this study, our aim is to predict the mean systolic blood pressure in some demographic domains of interest using the WC of NHANES participants as an auxiliary information which is likely measured with error

(Caballero 2005). The focus of our analysis is on 5588 participants. We build fifty domains ($m = 50$) with sample sizes ranging from 31 to 479, based on sex, five age categories (20-29, 30-39, 40-48, 50-59, and 60-84), and five race and ethnicity groups (Mexican American, Other Hispanic, White non-Hispanic, Black non-Hispanic and Other), and use the WC and systolic blood pressure as the values of the predictor and response variables. Figure 8 shows the mean systolic blood pressure versus the mean WC in 50 domains. It appears from Figure 8 that there is a non-linear relationship between these two variables. So, semi-parametric models such as P-spline models are good candidates to analyze this data set. In addition, since the WC is prone to measurement error, applying the proposed model (P-spline with $p = 1$ which considers ME of the variable WC) seems to be worthwhile. To compare the proposed model with the naive model, we also analyze this data set with ignoring measurement error. The estimated parameters (and standard errors using jackknife method) for the proposed model are $\hat{b}_0 = 35.92(20.41)$, $\hat{b}_1 = 0.88(0.20)$, $\hat{\sigma}^2_\eta = 234.33(13.07)$, $\hat{\sigma}^2_e = 249.27(23.13)$, $\hat{\sigma}^2_\nu = 5.00(6 \times 10^{-11})$, $\hat{\sigma}^2_\gamma = 0.13(0.05)$ and for the naive model are $\hat{b}_0 = 42.74(18.53)$, $\hat{b}_1 = 0.81(0.18)$, $\hat{\sigma}^2_e = 249.27(23.13)$, $\hat{\sigma}^2_\nu = 5.00(3 \times 10^{-11})$, $\hat{\sigma}^2_\gamma = 0.12(0.05)$. From the estimated parameter and corresponding standard error of variance of ME and comparing the test statistic with critical value 1.96, it is observed that the WC is measured with error. Furethermore, based on the obtained significant non-zero σ^2_γ, it is clear that there is a non-linear relationship between WC and systolic blood pressure. So, it seems that neither the non-linear relation nor the measurement error in WC, which cuases attenuation in estimate of the slope, can be ignored. It is also worth mentioning that the proposed model shows that the WC has a positive effect in predicting blood pressure which is supported by the literature (Lee, Bacha & Arslanian 2006; Choy et al. 2011; Duncan et al. 2013). The boxplots of PEB predictor of mean blood pressure for predefined domains, and their weighted and unweighted jackknife estimates of $MSPE$ for both models are presented in Figures 9a and 9b based on NHANES study. According to Figure 9a, it appears that both models behave similarly to predict mean blood pressure for predefined domains generally. We observe that mean blood pressure in men is higher than women fixing age, race and ethnicity. To study the effect of race and ethnicity, it is seen that mean blood pressure has approximately the same level in Mexican American, Other Hispanic, White non-Hispanic categories, while Black non-Hispanic and Other category have lower and White non-Hispanic group has higher level of blood pressure. Furthermore, based on the results, increasing age leads to a higher blood pressure, fixing the other two variables. In terms of $MSPE$ estimation, the naive jackknife estimation of $MSPE$ behave differently than the proposed model. In general, the weighted and unweighted estimators of $MSPE$ for the proposed model are smaller than the corresponding naive estimators of $MSPE$. Based on the simulation results, we can conclude that ignoring the ME may lead to wrong conclusions in terms of overestimation in jackknife $MSPE$ estimation of small area mean predictors.

[Figure 8 about here]

[Figure 9 about here]

5. Discussion

We have proposed a semi-parametric nested error regression model with functional ME in area-level covariate in a frequentist framework. According to the moment estimators in simulation part, it is observed that in the case of linear model as the true model, the regression coefficients of the proposed P-spline model are so close to the true values. Also, the estimate of variance of spline term is near zero in linear case while this estimate is non-zero for non-linear cases which indicates that the proposed model detects the existence of a non-linear relationship between the response and predictor variables. In particular, we have derived the PEB predictor of small area means and obtained the corresponding $MSPE$ of PEB predictor of small area means. We have also proposed jackknife estimators of the $MSPE$ of the PEB predictors. We have shown through a simulation that although the PEB predictor of small area means are very similar for the both proposed and naive methods, however, our proposed approach works very well in terms of jackknife $MSPE$ estimates of the PEB predictor of small area means compared to the naive model which ignores the ME in the covariate that causes serious overestimation due to the large value of cross-product terms involved in the MSPE of PEB predictor of small area means. We have also studied the effect of increasing ME in predictor variable for both models in different scenarios with considering different values for variance of ME. We have observed increasing ME causes attenuation in slope estimate in naive model unlike the proposed model.

Our proposed model is developed based on one covariate. An extension of our work to multiple covariates measured with error is simple, however, it will add much more unnecessary complexity to the model as one needs to define different spline terms for each covariate. One can also extend our approach to deal with non-normal random effects (Hui, Muller & Welsh 2020). As an extension of the proposed model, survey weights (Torabi 2011) can also be used in the estimation process in order to increase the efficiency of the PEB predictor of small area means. In this paper, we have assumed that the size of sample in the response and observed covariate is the same in each small area, however, one can extend our proposed approach and use multiple source of data with different sample sizes (Datta et al. 2018). One can extend our proposed model to generalized linear models (Torabi & Shokoohi 2015). One can also extend our proposed model and use the bootstrap and simple, unified, Monte-Carlo assisted (Sumca) methods as alternatives for $MSPE$ estimation of small area mean predictors (Jiang & Torabi 2020). These are some of the topics for future study.

Appendix

The supplementary materials provide R codes and corresponding "readme" files for the simulation and real application conducted in this paper.

References

BREIDT, F., CLAESKENS, G. & OPSOMER, J. (2005). Model-assisted estimation for complex surveys using penalised splines. *Biometrika* **92**, 831–846.

CABALLERO, B. (2005). *Encyclopedia of Human Nutrition.* Oxford: Elsevier Academic Press, 2nd edn.

CARROLL, RAYMOND J., D.R.L.A.S.C.M.C. (2006). *Measurement error in nonlinear models: a modern perspective.* Chapman and Hall/CRC, 2nd edn.

CHEN, S. & LAHIRI, P. (2002). On mean squared prediction error estimation in small area estimation problems. *In Proceedings of the Survey Research Methods Section* , 473–477.

CHOY, C.S., CHAN, W.Y., CHEN, T.L., SHIH, C.C., WU, L.C. & LIAO, C.C. (2011). Waist circumference and risk of elevated blood pressure in children: a cross–sectional study. *BMC Public Health* **11**, 1–7.

DATTA, G.S., RAO, J. & TORABI, M. (2010). Pseudo-empirical bayes estimation of small area means under a nested error linear regression model with functional measurement errors. *Journal of Statistical Planning and Inference* **140**, 2952–2962.

DATTA, G.S., TORABI, M., RAO, J. & LIU, B. (2018). Small area estimation with multiple covariates measured with errors: A nested error linear regression approach of combining multiple surveys. *Journal of Multivariate Analysis* **167**, 49–59.

DUNCAN, M.J., MOTA, J., VALE, S., SANTOS, M.P. & RIBEIRO, J.C. (2013). Associations between body mass index, waist circumference and body shape index with resting blood pressure in portuguese adolescents. *Annals of Human Biology* **40**, 163–167.

EILERS, P.H. & MARX, B.D. (1996). Flexible smoothing with b-splines and penalties. *Statistical science* , 89–102.

GHOSH, M. & SINHA, K. (2007). Empirical bayes estimation in finite population sampling under functional measurement error models. *Journal of statistical planning and inference* **137**, 2759–2773.

HASLETT, S. & WELSH, A. (2019). Eblup: Empirical best linear unbiased estimation. *Wiley StatsRef, Statistics Reference Online* .

HUI, F., MULLER, S. & WELSH, A. (2020). Random effects misspecification can have severe consequences for random effects inference in linear mixed models. *International Statistical Review* , 1–21.

HWANG, J. & KIM, D.H. (2015). Bayesian curve-fitting with radial basis functions under functional measurement error model. *Journal of the Korean Data and Information Science Society* **26**, 749–754.

HWANG, J. & KIM, D.H. (2016). Multivariable bayesian curve-fitting under functional measurement error model. *Journal of the Korean Data and Information Science Society* **27**, 1645–1651.

HWANG, J.S. & KIM, D.H. (2010). Semiparametric bayesian estimation under functional measurement error model. *Journal of the Korean Data and Information Science Society* **21**, 379–385.

JIANG, J. (2017). *Asymptotic analysis of mixed effects models: theory, applications, and open problems.* Chapman and Hall/CRC.

JIANG, J., LAHIRI, P., WAN, S.M. et al. (2002). A unified jackknife theory for empirical best prediction with m-estimation. *The Annals of Statistics* **30**, 1782–1810.

JIANG, J., NGUYEN, T. & RAO, J.S. (2010). Fence method for nonparametric small area estimation. *Survey Methodology* **36**, 3–11.

JIANG, J. & TORABI, M. (2020). Sumca: simple, unified, monte-carlo-assisted approach to second-order unbiased mean-squared prediction error estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* .

LEE, S., BACHA, F. & ARSLANIAN, S.A. (2006). Waist circumference, blood pressure, and lipid components of the metabolic syndrome. *The Journal of pediatrics* **149**, 809–816.

OPSOMER, J.D., CLAESKENS, G., RANALLI, M.G., KAUERMANN, G. & BREIDT, F. (2008). Non-parametric small area estimation using penalized spline regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**, 265–286.

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

PRATESI, M., RANALLI, M.G. & SALVATI, N. (2008). Semiparametric m-quantile regression for estimating the proportion of acidic lakes in 8-digit hucs of the northeastern us. *Environmetrics* **19**, 687–701.

PRATESI, M., RANALLI, M.G. & SALVATI, N. (2009). Nonparametric m-quantile regression using penalised splines. *Journal of Nonparametric Statistics* **21**, 287–304.

RAO, J. & MOLINA, I. (2015). *Small Area Estimation*. New Jersey: John Wiley & Sons, Inc, 2nd edn.

RAO, J.N., SINHA, S.K. & DUMITRESCU, L. (2014). Robust small area estimation under semi-parametric mixed models. *Canadian Journal of Statistics* **42**, 126–141.

RUPPERT, D. (2002). Selecting the number of knots for penalized splines. *Journal of computational and graphical statistics* **11**, 735–757.

RUPPERT, D., WAND, M.P. & CARROLL, R.J. (2003). *Semiparametric regression*. 12, Cambridge university press.

SALVATI, N., RANALLI, M. & PRATESI, M. (2011). Small area estimation of the mean using non-parametric m-quantile regression: a comparison when a linear mixed model does not hold. *Journal of Statistical Computation and Simulation* **81**, 945–964.

SHOKOOHI, F. & TORABI, M. (2018). Semi-parametric small-area estimation by combining time-series and cross-sectional data methods. *Australian & New Zealand Journal of Statistics* **60**, 323–342.

TORABI, M. (2011). Small area estimation using survey weights with functional measurement error in the covariate. *Australian & New Zealand Journal of Statistics* **53**, 141–155.

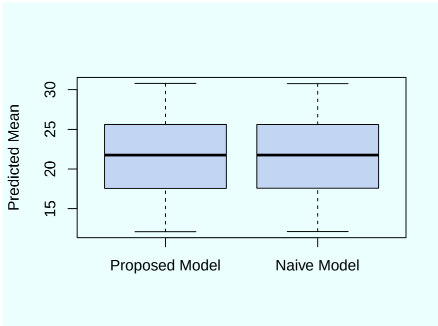
TORABI, M. & SHOKOOHI, F. (2015). Non-parametric generalized linear mixed models in small area estimation. *Canadian Journal of Statistics* **43**, 82–96.

TORKASHVAND, E., JAFARI JOZANI, M. & TORABI, M. (2015). Pseudo-empirical bayes estimation of small area means based on james–stein estimation in linear regression models with functional measurement error. *Canadian Journal of Statistics* **43**, 265–287.

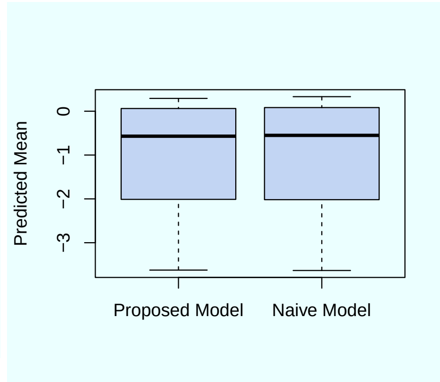
YBARRA, L.M. & LOHR, S.L. (2008). Small area estimation when auxiliary information is measured with error. *Biometrika* **95**, 919–931.

Table 1. Estimates (and standard errors) of the model parameters for the proposed model in the case of three forms for spline (linear, quadratic, exponential 1). True values in the case of linear model: $b_0 = b_1 = \sigma^2_\nu = \sigma^2_e = 1, \sigma^2_\eta = 2$; in the case of quadratic model: $b_0 = b_1 = 0.4, b_2 = -0.65, \sigma^2_\nu = 0.1, \sigma^2_e = 0.3, \sigma^2_\eta = 0.6$; in the case of exponential 1 model: $b_0 = b_1 = 1, b_2 = -0.7, \sigma^2_\nu = 0.1, \sigma^2_e = 0.3, \sigma^2_\eta = 0.6$.

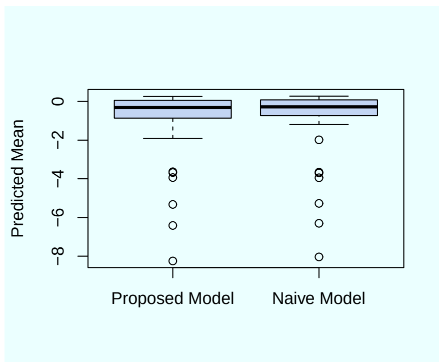
True model	b_0	b_1	σ^2_γ	σ^2_ν	σ^2_e	σ^2_η
Linear	0.94(0.89)	1.00(0.04)	$10^{-4}(5 \times 10^{-4})$	0.55(0.30)	1.00(0.13)	1.99(0.26)
Quadratic	-1.07(0.07)	0.13(0.06)	0.10(0.01)	$0.10(7 \times 10^{-13})$	0.31(0.04)	0.59(0.07)
Exponential 1	-0.97(0.08)	-0.73(0.07)	0.18(0.02)	$0.10(5 \times 10^{-13})$	0.30(0.04)	0.60(0.08)



(a)



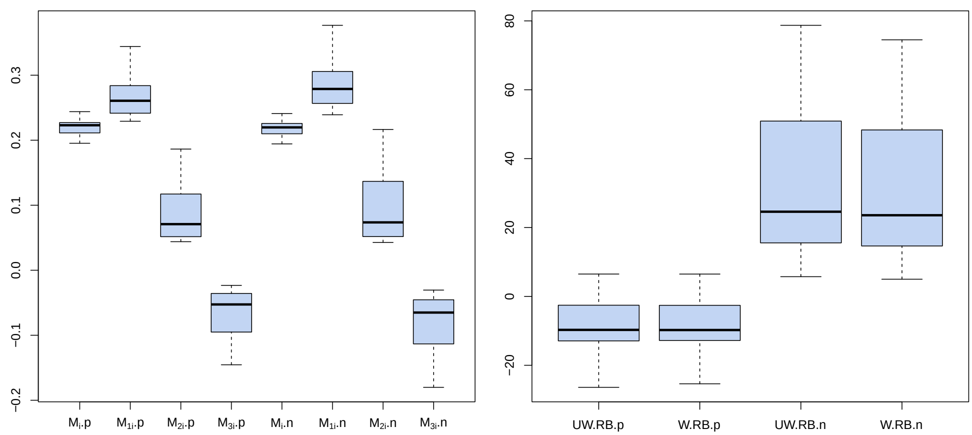
(b)



(c)

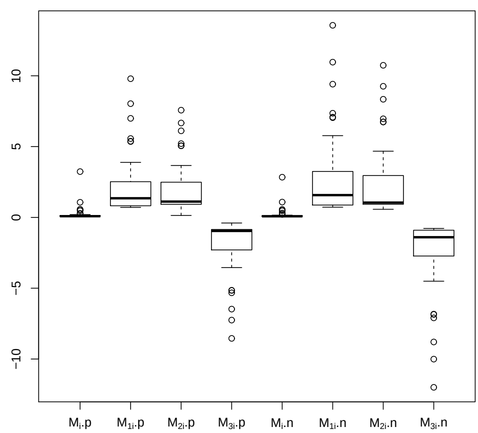
Figure 1. Boxplots of PEB predictions of small area means for the proposed and naive models in the case of (a) linear, (b) quadratic, (c) exponential 1.

450

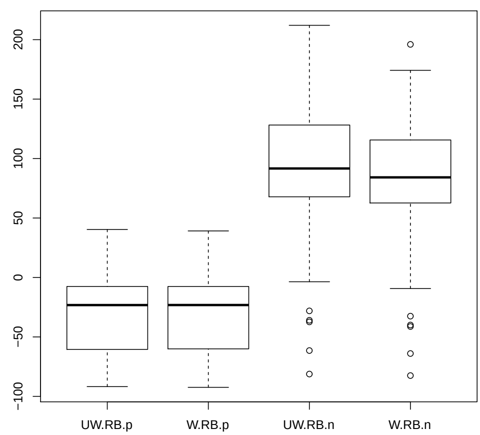


(a) (b)

Figure 2. (a) Boxplots of $EMSPE$ of PEB predictors and its components for the proposed model $(M_i.p, M_{1i}.p, M_{2i}.p, M_{3i}.p)$ and naive model $(M_i.n, M_{1i}.n, M_{2i}.n, M_{3i}.n)$ in the case of linear form for spline; (b) boxplots of percent relative bias of jackknife estimators of unweighted $(\mathrm{UW.RB}.p)$ and weighted $(\mathrm{W.RB}.p)$ $MSPE$ for the proposed model, unweighted $(\mathrm{UW.RB}.n)$ and weighted $(\mathrm{W.RB}.n)$ jackknife $MSPE$ estimation for the naive model in the case of linear form for spline.

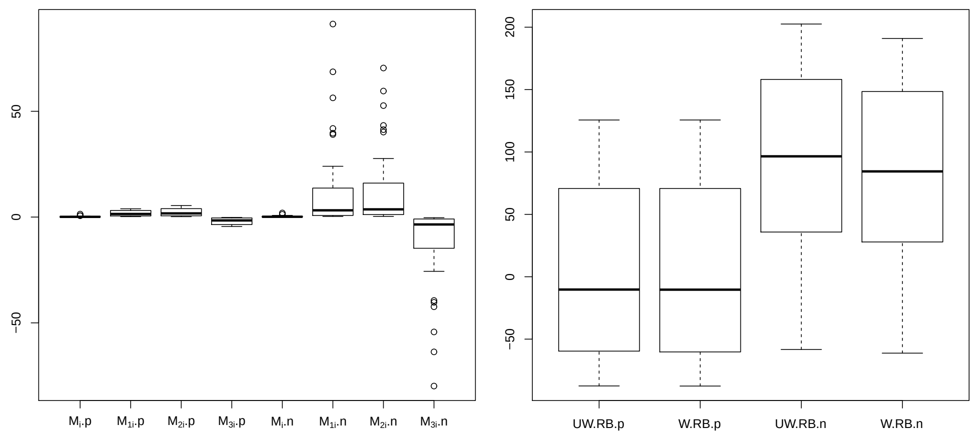


(a)



(b)

Figure 3. (a) Boxplots of $EMSPE$ of PEB predictors and its components for the proposed model $(M_i.p, M_{1i}.p, M_{2i}.p, M_{3i}.p)$ and naive model $(M_i.n, M_{1i}.n, M_{2i}.n, M_{3i}.n)$ in the case of quadratic form for spline; (b) boxplots of percent relative bias of jackknife estimators of unweighted $(\mathrm{UW.RB}.p)$ and weighted $(\mathrm{W.RB}.p)$ $MSPE$ for the proposed model, unweighted $(\mathrm{UW.RB}.n)$ and weighted $(\mathrm{W.RB}.n)$ jackknife $MSPE$ estimation for the naive model in the case of quadratic form for spline.

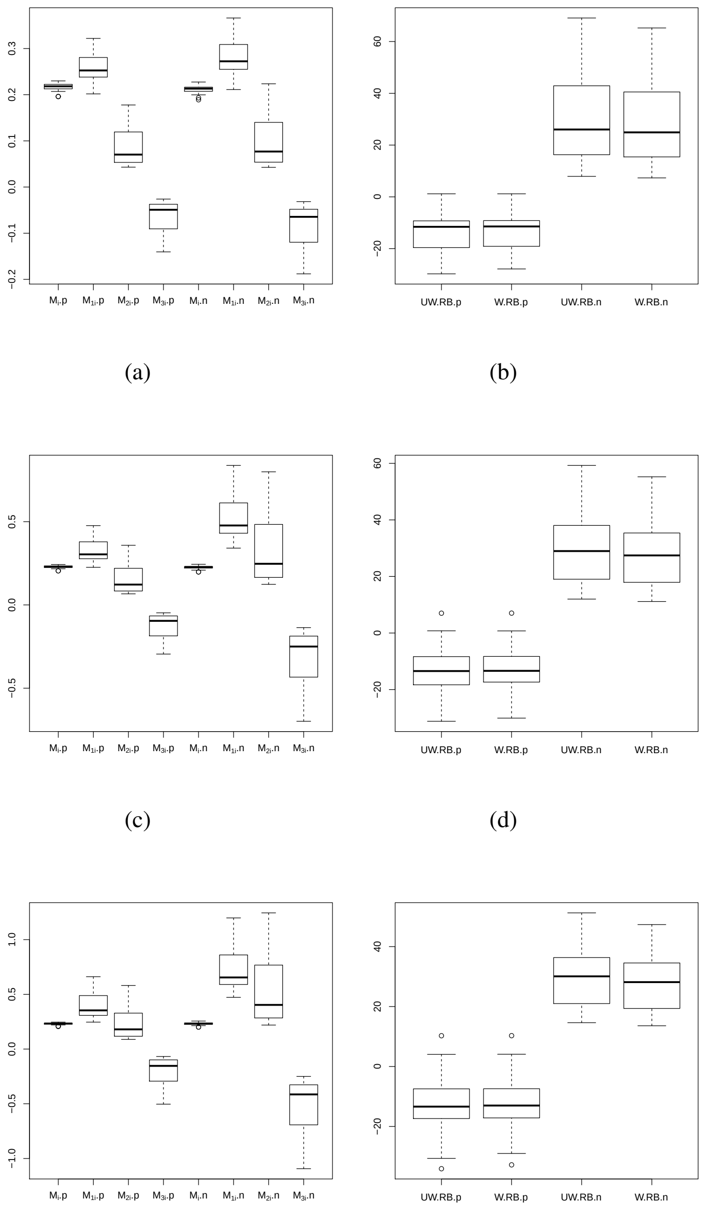


(a) (b)

Figure 4. (a) Boxplots of $EMSPE$ of PEB predictors and its components for the proposed model $(M_i.p, M_{1i}.p, M_{2i}.p, M_{3i}.p)$ and naive model $(M_i.n, M_{1i}.n, M_{2i}.n, M_{3i}.n)$ in the case of exponential 1 form for spline; (b) boxplots of percent relative bias of jackknife estimators of unweighted $(\mathrm{UW.RB.}p)$ and weighted $(\mathrm{W.RB.}p)$ $MSPE$ for the proposed model, unweighted $(\mathrm{UW.RB.}n)$ and weighted $(\mathrm{W.RB.}n)$ jackknife $MSPE$ estimation for the naive model in the case of exponential 1 form for spline.

Table 2. Estimates (and standard errors) of the model parameters for the proposed model in the case of three forms for spline (linear, quadratic, exponential 1). True values in the case of linear model: $b_0 = b_1 = \sigma_\nu^2 = \sigma_e^2 = 1$; in the case of quadratic model: $b_0 = b_1 = 0.4, b_2 = -0.65, \sigma_\nu^2 = 0.1, \sigma_e^2 = 0.3$; in the case of exponential 1 model: $b_0 = b_1 = 1, b_2 = -0.7, \sigma_\nu^2 = 0.1, \sigma_e^2 = 0.3$.

True model	True value of σ_η^2	b_0	b_1	σ_γ^2	σ_ν^2	σ_e^2	σ_η^2
Linear	1	1.00 (0.20)	1.01 (0.14)	$2 \times 10^{-3}(6 \times 10^{-3})$	0.68 (0.26)	1.00 (0.13)	1.01 (0.12)
	3.5	1.00 (0.24)	1.05 (0.22)	$10^{-3}(4 \times 10^{-3})$	0.37 (0.36)	1.00 (0.13)	3.52 (0.43)
	5	0.99 (0.28)	1.08 (0.30)	$10^{-3}(4 \times 10^{-3})$	0.29 (0.36)	1.00 (0.13)	5.03 (0.62)
Quadratic	0.2	-1.07 (0.06)	0.12 (0.04)	0.10 (0.01)	$0.10(3 \times 10^{-3})$	0.30 (0.03)	0.19 (0.02)
	1	-1.07 (0.06)	0.12 (0.06)	0.12(0.02)	$0.10(7 \times 10^{-13})$	0.30 (0.04)	0.93 (0.12)
	1.5	-1.06 (0.06)	0.11 (0.07)	0.14 (0.03)	$0.10(6 \times 10^{-13})$	0.30 (0.03)	1.45 (0.16)
Exponential 1	0.2	-0.97 (0.07)	-0.73 (0.05)	0.17 (0.01)	$0.10(5 \times 10^{-13})$	0.30 (0.03)	0.19 (0.02)
	1	-0.97 (0.09)	-0.74 (0.09)	0.19 (0.03)	$0.10(6 \times 10^{-13})$	0.30 (0.03)	1.00 (0.13)
	1.5	-0.97 (0.09)	-0.75 (0.10)	0.18 (0.03)	$0.10(6 \times 10^{-13})$	0.30 (0.03)	1.48 (0.18)



(a) (b)

(c) (d)

(e) (f)

Figure 5. Boxplots of $EMSPE$ of PEB predictors and its components for the proposed model $(M_i.p, M_{1i}.p, M_{2i}.p, M_{3i}.p)$ and naive model $(M_i.n, M_{1i}.n, M_{2i}.n, M_{3i}.n)$ in the case of linear form for spline with (a) $\sigma_\eta^2 = 1.00$, (c) $\sigma_\eta^2 = 3.50$, (e) $\sigma_\eta^2 = 5.00$; boxplots of percent relative bias of jackknife estimators of unweighted $(\mathrm{UW.RB.}p)$ and weighted $(\mathrm{W.RB.}p)$ $MSPE$ for the proposed model, unweighted $(\mathrm{UW.RB.}n)$ and weighted $(\mathrm{W.RB.}n)$ jackknife $MSPE$ estimation for the naive model in the case of linear form for spline with (b) $\sigma_\eta^2 = 1.00$, (d) $\sigma_\eta^2 = 3.50$, (f) $\sigma_\eta^2 = 5.00$.

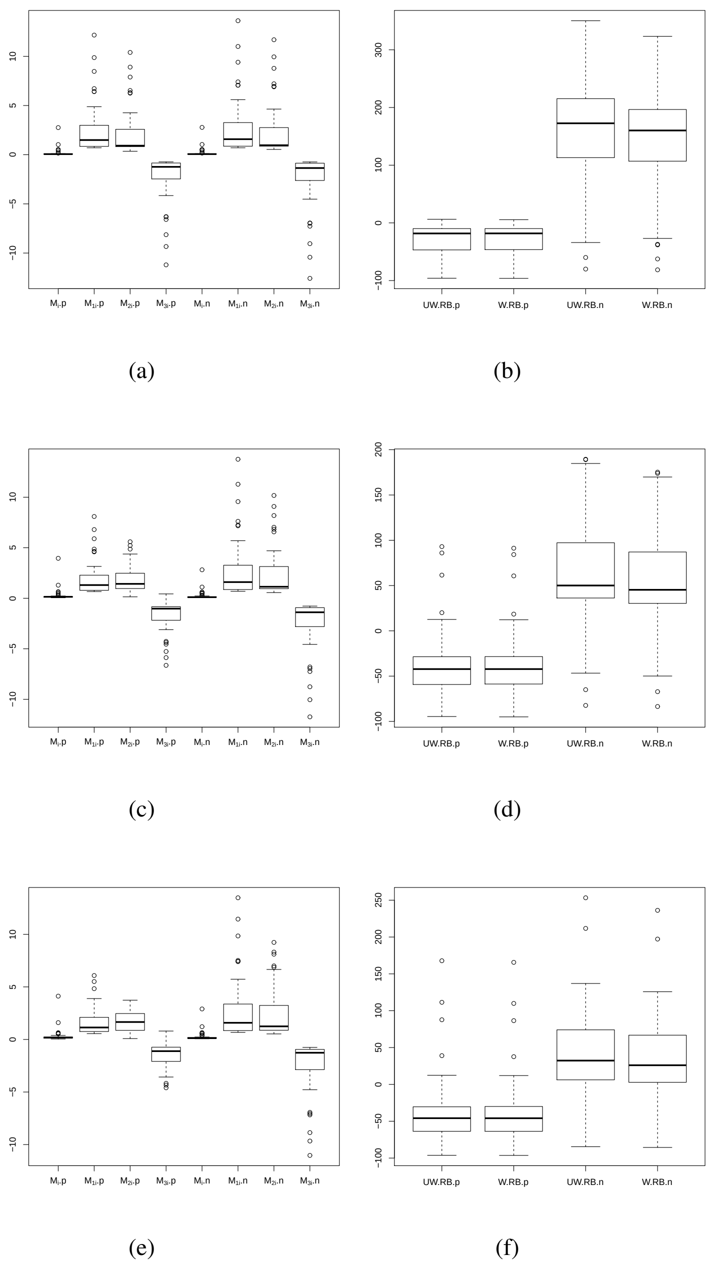


Figure 6. Boxplots of $EMSPE$ of PEB predictors and its components for the proposed model $(M_i.p, M_{1i}.p, M_{2i}.p, M_{3i}.p)$ and naive model $(M_i.n, M_{1i}.n, M_{2i}.n, M_{3i}.n)$ in the case of quadratic form for spline with (a) $\sigma_\eta^2 = 0.20$, (c) $\sigma_\eta^2 = 1.00$, (e) $\sigma_\eta^2 = 1.5$; boxplots of percent relative bias of jackknife estimators of unweighted $(\mathrm{UW.RB}.p)$ and weighted $(\mathrm{W.RB}.p)$ $MSPE$ for the proposed model, unweighted $(\mathrm{UW.RB}.n)$ and weighted $(\mathrm{W.RB}.n)$ jackknife $MSPE$ estimation for the naive model in the case of quadratic form for spline with (b) $\sigma_\eta^2 = 0.20$, (d) $\sigma_\eta^2 = 1.00$, (f) $\sigma_\eta^2 = 1.50$.

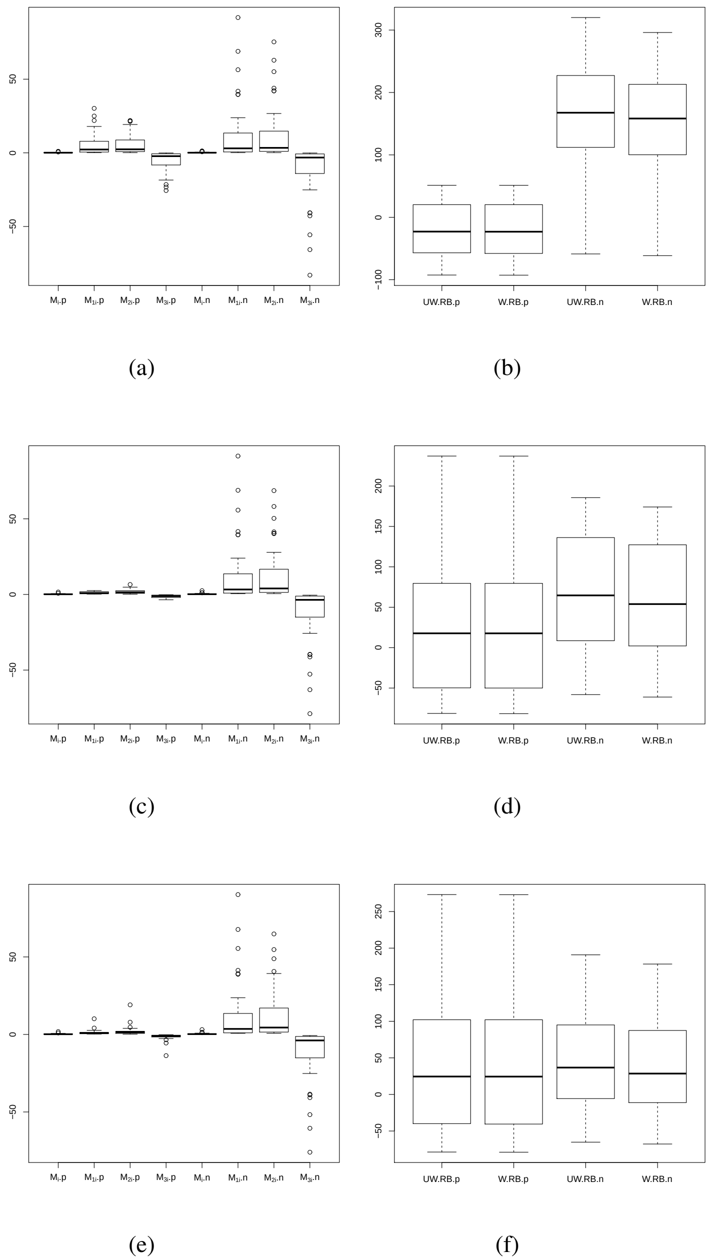


Figure 7. Boxplots of $EMSPE$ of PEB predictors and its components for the proposed model $(M_i.p, M_{1i}.p, M_{2i}.p, M_{3i}.p)$ and naive model $(M_i.n, M_{1i}.n, M_{2i}.n, M_{3i}.n)$ in the case of exponential 1 form for spline with (a) $\sigma_\eta^2 = 0.20$, (c) $\sigma_\eta^2 = 1.00$, (e) $\sigma_\eta^2 = 1.50$; boxplots of percent relative bias of jackknife estimators of unweighted $(\mathrm{UW.RB.}p)$ and weighted $(\mathrm{W.RB.}p)$ $MSPE$ for the proposed model, unweighted $(\mathrm{UW.RB.}n)$ and weighted $(\mathrm{W.RB.}n)$ jackknife $MSPE$ estimation for the naive model in the case of exponential 1 form for spline with (b) $\sigma_\eta^2 = 0.20$, (d) $\sigma_\eta^2 = 1.00$, (f) $\sigma_\eta^2 = 1.50$.

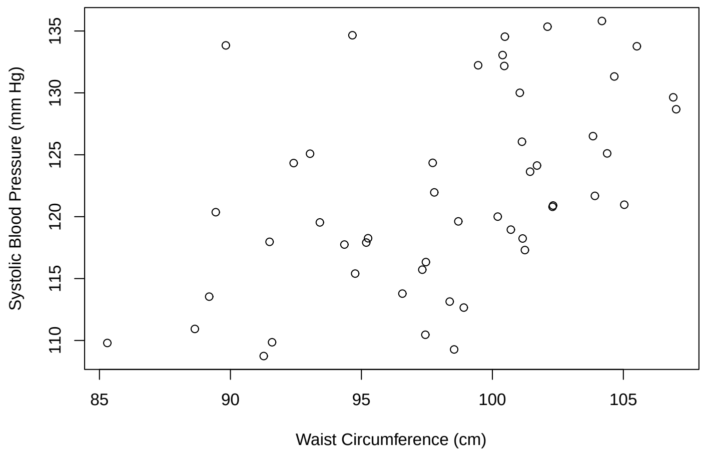
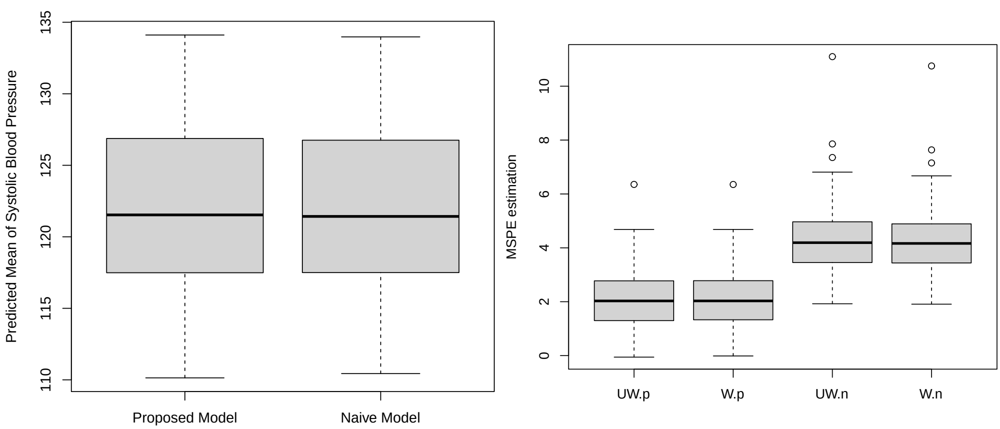


Figure 8. Average systolic blood pressure versus average waist circumference for some predefined groups (sex-age-race and ethnicity) based on US NHANES 2013– 2014 data.



(b)

(a)

Figure 9. (a) Boxplots of PEB predictors of small area blood pressure means for the proposed and naive models; (b) boxplots of unweighted (UW.p) and weighted (W.p) jackknife estimates of $MSPE$ of small area blood pressure mean predictors for the proposed model; boxplots of unweighted (UW.n) and weighted (W.n) jackknife estimates of $MSPE$ of small area blood pressure mean predictors for the naive model. The boxplots in parts (a) and (b) are for some predefined groups (sex-age-race and ethnicity) based on US NHANES 2013– 2014 data.

Appendix A Supplementary Table

Table A1. Variable definitions

Variable	Definition
m	Number of areas
N_i	Population size in i-th small area
n_i	Sample size in i-th small area
n_T	Total sample size
y_{ij}	Response variable for j-th unit at the i-th small area
$\boldsymbol{y}_i$	Vector of response variable in i-th area
$\boldsymbol{y}$	Vector of response variable
x_i	True value of the covariate in i-th area
$\boldsymbol{x}_i$	Vector of true value of the covariates of fixed part of the model
$\boldsymbol{X}$	Matrix of true value of the covariates of fixed part of the model
$\boldsymbol{z}_i$	Vector of true value of the covariates of P-spline part of the model
$\boldsymbol{Z}$	Matrix of true value of the covariates of P-spline part of the model
w_{ij}	Observed value of the covariate for j-th unit at the i-th small area
$\boldsymbol{\nu} = (\nu_1, ..., \nu_m)$	Vector of area-level random effects
$\boldsymbol{d}_{ij}$	Vector that shows (i, j)-th sample belongs to which area
$\boldsymbol{D}$	Matrix that shows each sample belongs to which area
e_{ij}	Random error for j-th unit at the i-th small area
$\boldsymbol{e}_i$	Vector of random errors in i-th area
$\boldsymbol{e}$	Matrix of random errors
η_{ij}	Measurement error for j-th unit at the i-th small area
$\boldsymbol{b} = (b_0, \ldots, b_p)$	Vector of regression coefficients of fixed part of the model
$\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_k)$	Vector of regression coefficients of P-spline part of the model
p	Degree of spline
$(\tau_1, \ldots, \tau_k)$	Set of knots
σ_ν^2	Variance of the area-level random effects
σ_γ^2	Variance of the random effects of P-spline part of the model
σ_e^2	Variance of the random error
σ_η^2	Variance of the measurement error
$\boldsymbol{V}$	Variance-covariance matrix of $\boldsymbol{Y}$
$\boldsymbol{\Sigma}_\nu$	Variance-covariance matrix of $\boldsymbol{\nu}$
$\boldsymbol{\Sigma}_\gamma$	Variance-covariance matrix of $\boldsymbol{\gamma}$
$\boldsymbol{\Sigma}_e$	Variance-covariance matrix of $\boldsymbol{e}$
$\boldsymbol{\Sigma}_\eta$	Variance-covariance matrix of $\boldsymbol{\eta}$
θ_i	Mean response variable of i-th small area
θ_i^B	Best predictor of i-th small area
θ_i^{PB}	Pseudo-best predictor of i-th area
θ_i^{PEB}	Pseudo-empirical best predictor of i-th small area
$\tilde{x}_i$	Estimate of x_i when all the parameters are known
$\hat{x}_i$	Estimate of x_i when all the parameters are estimated
$\mathbf{h}_i = (1, \bar{W}_i, \ldots, \bar{W}_i^p)^\top$	Vector of the observed value of the covariates mean
R	Number of simulation runs in simulation study
$mspe_J$	Unweighted jackknife mean squared prediction error estimator
$mspe_{JW}$	Weighted jackknife mean squared prediction error estimator
RB	Relative Bias