Journal of the Korean Data & Information Science Society
2010, **21**(2), 379–385

Semiparametric Bayesian estimation under functional measurement error model

Jin Seub Hwang[1] · Dal Ho Kim[2]

[12]Department of Statistics, Kyungpook National University

Received 15 February 2010, revised 10 March 2010, accepted 20 March 2010

Abstract

This paper considers Bayesian approach to modeling a flexible regression function under functional measurement error model. The regression function is modeled based on semiparametric regression with penalized splines. Model fitting and parameter estimation are carried out in a hierarchical Bayesian framework using Markov chain Monte Carlo methodology. Their performances are compared with those of the estimators under functional measurement error model without semiparametric component.

Keywords: Functional measurement error, Gibbs sampling, hierarchical Bayes, penalized spline, semiparametric.

1. Introduction

Small area estimation has received considerable attention due to growing demand for reliable small area statistics in both public and private sectors. Rao (2003) gives a comprehensive account of model-based methods that lead to efficient estimators of small area means when the area-specific sample sizes are small.

Ghosh and Meeden (1986) considered EB estimation in a stratified finite population context using a simple one-way ANOVA model. The results can be extended by inclusion of covariates, and such procedures have been discussed in Ghosh and Meeden (1996). Often, however, it is not possible to obtain exact measurements of covariates. Ghosh and Sinha (2004), abbreviated GS, assumed that the covariates are measured with error and non-stochastic. This is the so-called functional measurement error model.

Semiparametric regression methods have not been used in small area estimation contexts until recently. This was mainly due to methodological difficulties in combining the different smoothing techniques with the estimation tools generally used in small area estimation. The pioneering contribution in this regard is the work by Opsomer *et al.* (2008) in which they combined small area random effects with a smooth.

The objective of this article is to develop efficient estimators of small area means by using flexible smoothing of non-linear pattern with functional measurement error model. In doing

[1] Ph.D, Department of Statistics, Kyungpook National University, Daegu 702-701, Korea.
[2] Corresponding author: Professor, Department of Statistics, Kyungpook National University, Daegu 702-701, Korea. E-mail: dalkim@knu.ac.kr

so, we have modeled the small area means using penalized spline (Eilers and Marx, 1996) which is a commonly used but powerful function estimation tool in nonparametric inference. We have used truncated polynomial basis functions with varying degrees and number of knots, although other types of basis functions like B-splines or thinplate splines can also be used. For our semiparametric model, the analysis has been carried out using a hierarchical Bayesian (HB) approach. To sum up, we develop HB procedures for semiparametric small area estimation under functional measurement error model. Following the general convention, we have placed the knots on a grid of equally spaced sample quantiles of the independent variables.

The model specification are given in Section 2. In Section 3, we have established the propriety of the posteriors, and have discussed the Markov chain Monte Carlo (MCMC) implementation of the proposed hierarchical Bayes procedure. Analysis of some real-life data is undertaken in Section 4. Finally, we present a discussion of the results in Section 5. The proofs of certain technical results are deferred to the Appendix.

2. Model specification

Suppose there are m strata labelled $1, \cdots, m$ and let N_i denote the known population size for the i th stratum. We denote by y_{ij} the response of the j th unit in the i th stratum $(j = 1, \cdots, N_i; i = 1, \cdots, m)$. A sample of size n_i is drawn from the i th stratum $(\sum_{i=1}^{m} n_i = n_t)$. We consider the superpopulation model

$$y_{ij} = \boldsymbol{x_i^T b} + \boldsymbol{z_i^T \gamma} + u_i + e_{ij} \quad (j = 1, \cdots, N_i; i = 1, \cdots, m), \tag{2.1}$$

$$X_{ij} = x_i + \eta_{ij} \quad (j = 1, \cdots, N_i; i = 1, \cdots, m). \tag{2.2}$$

where $\boldsymbol{x}_i = (1, x_i)^T$, $\boldsymbol{z}_i = \{(x_i - \tau_1)_+, \cdots, (x_i - \tau_k)_+\}^T$, $\boldsymbol{b} = (b_0, b_1)^T$, and $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_k)^T$. Here k is the number of knots. It is assumed that the u_i, e_{ij} and η_{ij} are mutually independent with $u_i \sim N(0, \sigma_u^2)$, $e_{ij} \sim N(0, \sigma_e^2)$ and $\eta_{ij} \sim N(0, \sigma_\eta^2)$. The available data consist of (y_{ij}, X_{ij}). An alternative way of expressing the same is

$$y_{ij} = \theta_i + e_{ij}; \theta_i = \boldsymbol{x_i^T b} + \boldsymbol{z_i^T \gamma} + u_i, \quad (j = 1, \cdots, N_i; i = 1, \cdots, m) \tag{2.3}$$

Here the goal is to estimate the small area means $\boldsymbol{\theta}$.

3. Hierachical Bayesian inference

We consider a hierarchical Bayesian framework to predict the small area means $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_m)$. Using expression (3), we begin with the following HB model:

Stage 1. $y_{ij} = \theta_i + e_{ij}$ $(j = 1, \cdots, n_i; i = 1, \cdots, m)$ where $e_{ij} \overset{iid}{\sim} N(0, \sigma_e^2)$

Stage 2. $\theta_i = \boldsymbol{x_i^T b} + \boldsymbol{z_i^T \gamma} + u_i$ $(i = 1, \cdots, m)$ where $u_i \overset{iid}{\sim} N(0, \sigma_u^2)$
$X_{ij} = x_i + \eta_{ij}$ $(j = 1, \cdots, n_i; i = 1, \cdots, m)$ where $\eta_{ij} \overset{iid}{\sim} N(0, \sigma_\eta^2)$

Stage 3. $\boldsymbol{\gamma} \sim N(0, \sigma_\gamma^2 I)$

Stage 4. $b_0, b_{1,}\sigma_e^2, \sigma_u^2, \sigma_\eta^2$ and σ_γ^2 are mutually independent with b_0 and b_1 i.i.d. $uniform(-\infty,\infty)$; $(\sigma_e^2)^{-1} \sim G(a_e, b_e)$, $(\sigma_u^2)^{-1} \sim G(a_u, b_u)$, $(\sigma_\eta^2)^{-1} \sim G(a_\eta, b_\eta)$, $(\sigma_\gamma^2)^{-1} \sim G(a_\gamma, b_\gamma)$, where $G(\alpha,\beta)$ denotes an gamma distribution with shape parameter α and rate parameter β having the expression $f(x) \propto x^{\alpha-1}\exp(-\beta x), x \geq 0$.

First check the propriety of the posterior under the given prior. By the conditional independence properties, we can factorize the full posterior as

$$\begin{aligned} &\left[\boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{\gamma}, \sigma_e^2, \sigma_u^2, \sigma_\eta^2, \sigma_\gamma^2 | \boldsymbol{X}, \boldsymbol{y}\right] \\ \propto &\left[\boldsymbol{y}|\boldsymbol{\theta}, \sigma_e^2\right] \left[\boldsymbol{\theta}|\boldsymbol{b}\boldsymbol{,}\boldsymbol{\gamma}\boldsymbol{,}\sigma_u^2, \boldsymbol{X}\right] \left[\boldsymbol{X}|\sigma_\eta^2\right] \left[\boldsymbol{\gamma}|\sigma_\gamma^2\right] \left[\boldsymbol{b}\right] \left[\sigma_e^2\right] \left[\sigma_u^2\right] \left[\sigma_\eta^2\right] \left[\sigma_\gamma^2\right] \end{aligned} \tag{3.1}$$

The proof of the propriety of the posterior is deferred to the Appendix.

The implementation of the Bayesian procedure is greatly facilitated by the Markov chain Monte Carlo numerical integration technique, in particular the Gibbs sampler. This requires generating samples from the full conditions of each of $\boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{\gamma}, \sigma_e^2, \sigma_u^2, \sigma_\eta^2$ and σ_γ^2 given the remaining parameters and the data. The Gibbs sampling analysis is based on the following full conditional distribution:

(i) $\left[\theta_i|\boldsymbol{b}, \boldsymbol{\gamma}, \sigma_e^2, \sigma_u^2, \sigma_\gamma^2, \sigma_\eta^2, \boldsymbol{X}, \boldsymbol{y}\right] \overset{iid}{\sim} N\left[(1-C_i)\,\bar{y}_i + C_i\left(\boldsymbol{x_i^T b} + \boldsymbol{z_i^T \gamma}\right), \sigma_e^2/n_i\,(1-C_i)\right]$
where $C_i = \sigma_e^2/\left(\sigma_e^2 + n_i\sigma_u^2\right)$

(ii) $\left[\boldsymbol{b}|\boldsymbol{\theta}, \boldsymbol{\gamma}, \sigma_e^2, \sigma_u^2, \sigma_\gamma^2, \sigma_\eta^2, \boldsymbol{X}, \boldsymbol{y}\right] \sim N\left[\left(\boldsymbol{X_*^T X_*}\right)^{-1}\boldsymbol{X_*^T w}, \sigma_u^2\left(\boldsymbol{X_*^T X_*}\right)^{-1}\right]$
where $\boldsymbol{X_*} = \left(\boldsymbol{x_1^T}, \cdots, \boldsymbol{x_m^T}\right)^T$, $\boldsymbol{w} = (w_1, \cdots, w_m)^T, w_i = \theta_i - \boldsymbol{z_i^T \gamma}$

(iii) $\left[\boldsymbol{\gamma}|\boldsymbol{\theta}, \boldsymbol{b}, \sigma_e^2, \sigma_u^2, \sigma_\gamma^2, \sigma_\eta^2, \boldsymbol{X}, \boldsymbol{y}\right] \sim N\left[\left(\frac{\boldsymbol{Z_*^T Z_*}}{\sigma_u^2} + \frac{I}{\sigma_\gamma^2}\right)^{-1}\frac{\boldsymbol{Z_*^T}}{\sigma_u^2}\boldsymbol{t}, \left(\frac{\boldsymbol{Z_*^T Z_*}}{\sigma_u^2} + \frac{I}{\sigma_\gamma^2}\right)^{-1}\right]$

where $\boldsymbol{Z_*} = \begin{pmatrix} (x_1-\tau_1)_+ & \cdots & (x_1-\tau_k)_+ \\ \vdots & \vdots & \vdots \\ (x_m-\tau_1)_+ & \cdots & (x_m-\tau_k)_+ \end{pmatrix}$, $\boldsymbol{t} = (t_1, \cdots, t_m)^T, t_i = \theta_i - \boldsymbol{x_i^T b}$

(iv) $\left[\sigma_e^{-2}|\boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{\gamma}\boldsymbol{,}\sigma_u^2, \sigma_\gamma^2, \sigma_\eta^2, \boldsymbol{X}\boldsymbol{,}\boldsymbol{y}\right] \sim G\left[\frac{n_t}{2} + a_e, \frac{1}{2}\sum_{i=1}^m \sum_{j=1}^{n_i}\left(y_{ij} - \theta_i\right)^2 + b_e\right]$

(v) $\left[\sigma_u^{-2}|\boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{\gamma}\boldsymbol{,}\sigma_e^2, \sigma_\gamma^2, \sigma_\eta^2, \boldsymbol{X}\boldsymbol{,}\boldsymbol{y}\right] \sim G\left[\frac{m}{2} + a_u, \frac{1}{2}\sum_{i=1}^m\left(\theta_i - \boldsymbol{x_i^T b} - \boldsymbol{z_i^T \gamma}\right)^2 + b_u\right]$

(vi) $\left[\sigma_\eta^{-2}|\boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{\gamma}\boldsymbol{,}\sigma_e^2, \sigma_\gamma^2, \sigma_u^2, \boldsymbol{X}\boldsymbol{,}\boldsymbol{y}\right] \sim G\left[\frac{n_t}{2} + a_\eta, \frac{1}{2}\sum_{i=1}^m \sum_{j=1}^{n_i}\left(X_{ij} - x_i\right)^2 + b_\eta\right]$

(vii) $\left[\sigma_\gamma^{-2}|\boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{\gamma}\boldsymbol{,}\sigma_e^2, \sigma_u^2, \sigma_\eta^2, \boldsymbol{X}\boldsymbol{,}\boldsymbol{y}\right] \sim G\left[\frac{k}{2} + a_\gamma, \frac{1}{2}\boldsymbol{\gamma^T \gamma} + b_\gamma\right]$

We generate several sets of samples from the above full conditional distributions. After burning out the first half, we use the averaging principle and take the average of the HB estimates over all the remaining sets to obtain the final HB estimates. Also, we are replace x_i by $\bar{X}_i = n_i^{-1} \sum_{j=1}^{n_i} X_{ij}$. The HB estimators for small area means is approximated as:

$$\begin{aligned} E\left(\theta_i|\boldsymbol{X},\boldsymbol{y}\right) &= E\left[E\left(\theta_i|\boldsymbol{b},\boldsymbol{\gamma},\sigma_e^2,\sigma_u^2,\sigma_\gamma^2,\sigma_\eta^2,\boldsymbol{X},\boldsymbol{y}\right)\right] \qquad (3.2)\\ &\simeq \left(d^{-1}\right) \sum_{r=d+1}^{2d} \left[\left(1-C_i^{(r)}\right)\bar{y}_i + C_i^{(r)}\left(\boldsymbol{x_i^T b^{(r)}} + \boldsymbol{z_i^T \gamma^{(r)}}\right)\right] \end{aligned}$$

$$\begin{aligned} V\left(\theta_i|\boldsymbol{X},\boldsymbol{y}\right) &= E\left[V\left(\theta_i|\boldsymbol{b},\boldsymbol{\gamma},\sigma_e^2,\sigma_u^2,\sigma_\gamma^2,\sigma_\eta^2,\boldsymbol{X},\boldsymbol{y}\right)\right] + V\left[E\left(\theta_i|\boldsymbol{b},\boldsymbol{\gamma},\sigma_e^2,\sigma_u^2,\sigma_\gamma^2,\sigma_\eta^2,\boldsymbol{X},\boldsymbol{y}\right)\right]\\ &\simeq \left(d^{-1}\right) \sum_{r=1}^{2d} \left(\frac{\sigma_e^{2(r)}}{n_i}(1-C_i^{(r)})\right) \qquad (3.3)\\ &+ \left(d^{-1}\right) \sum_{r=d+1}^{2d} \left[1-C_i^{(r)}\bar{y}_i + C_i^{(r)}\left(\boldsymbol{x_i^T b^{(r)}} + \boldsymbol{z_i^T \gamma^{(r)}}\right)\right]^2\\ &- \left[E(\theta_i|\boldsymbol{X},\boldsymbol{y})\right]^2 . \end{aligned}$$

We use these results in next section for finding the posterior means and variances of the θ_i's ($i = 1, ..., m$).

4. Data analysis

We conducted a data analysis to compare the performance of the proposed HB estimators and the HB estimators without semiparametric component. We use the data used by Battese *et al.* (1988), hereafter BHF, for analysis. Knowledge of the area under different crops is important to the US Department of Agriculture (USDA). Sample surveys have designed to estimate crop areas for large regions, such as crop-reporting districts, individual states, and the USA as a whole. Predicting crop areas for small areas such as counties has generally not been attempted, due to a lack of availability of data from farm surveys for these areas. The use of satellite data in association with farm-land survey observations has been the subject of considerable research over the years. In their paper, BHF considered data for 12 counties in Iowa, obtained from the 1978 June Enumerative Survey of the USDA as well as from the satellite LANDSAT during the 1978 growing season. The purpose was to predict the area under soya bean and corn in these counties. BHF developed a variance components model for small area estimation and they provided analysis of the soya bean data (reported by farmers) using two covariates, corn and soya bean (reported by satellite). The actual data are provided in BHF (1988). We consider prediction of corn data using corn pixels only as covariates.

We ran a Gibbs chain of size 10,000 with a burn-in of the first 5000. Using the equation (5) and (6), we estimated of small area means and standard error. Figure 1 shows the scatter plot of (y_{ij}, X_{ij}). The results are reported in Table 4.1. It can be seen that the proposal model with one knot is much better than the GS model in view of standard error. Sometimes experience on the subject matter may be a guiding force in placing the knots the "optimum" locations where a sharp change in the curve pattern can be expected.

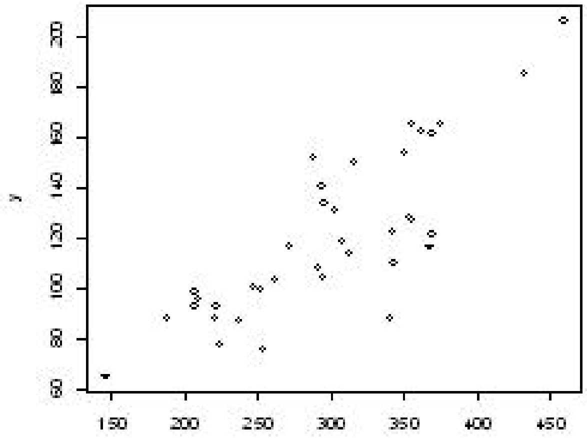

Figure 4.1 Scatter plot for BHF data

Table 4.1 The sample sizes, estimates and s.e. for the 12 counties

i	n_i	GS		1 knot		2 knots		3 knots	
		Est	SE	Est	SE	Est	SE	Est	SE
1	1	150.01	9.80	151.08	9.82	151.23	10.98	150.87	10.08
2	1	85.09	11.91	94.54	12.00	90.71	14.87	93.77	13.23
3	1	101.83	8.55	106.14	7.41	104.72	7.79	104.83	8.44
4	2	159.57	11.65	161.10	12.21	162.46	14.53	160.55	13.28
5	3	136.95	7.58	137.37	6.91	135.97	9.48	137.44	8.14
6	3	90.57	10.58	98.38	9.92	95.33	11.52	97.60	10.75
7	3	116.98	6.60	116.89	5.67	118.06	6.36	116.12	7.42
8	3	141.24	8.00	141.87	7.54	141.02	9.13	141.72	8.33
9	4	106.72	7.72	109.83	6.10	108.98	7.36	108.99	7.05
10	5	114.11	6.63	114.81	5.38	115.48	6.27	113.94	6.79
11	5	125.28	6.57	125.12	5.66	125.38	7.07	124.32	7.04
12	6	114.95	6.52	115.49	5.27	116.24	6.14	114.74	6.74

5. Discussion

In this paper, we develop HB procedures for semiparametric small area estimation under functional measurement error model with fixed knots. We noted that the result depends on the number of knots as well as the location of knots. We will pursue further the semiparametric Bayesian estimation with random knots. Also, we will consider semiparametric Bayesian estimators under structural measurement error models where the covariates are measured with error and stochastic.

Appendix A: Proof of posterior propriety

The basic parameter space is $\boldsymbol{\Omega} = \left\{\boldsymbol{\theta},\boldsymbol{b},\boldsymbol{\gamma},\sigma_e^2,\sigma_u^2,\sigma_\eta^2,\sigma_\gamma^2\right\}$. Let

$$\begin{aligned} I &= \int \cdots \int p(\Omega|\boldsymbol{y},\boldsymbol{X})d\boldsymbol{\Omega} \qquad \text{(A.1)} \\ &= \int \cdots \int \left[\boldsymbol{y}|\boldsymbol{\theta},\sigma_e^2\right]\left[\boldsymbol{\theta}|\boldsymbol{b},\boldsymbol{\gamma},\boldsymbol{\sigma_u^2},\boldsymbol{X}\right]\left[\boldsymbol{X}|\sigma_\eta^2\right]\left[\gamma|\sigma_\gamma^2\right][\boldsymbol{b}]\left[\sigma_e^2\right]\left[\sigma_u^2\right]\left[\sigma_\eta^2\right]\left[\sigma_\gamma^2\right]d\boldsymbol{\Omega} \end{aligned}$$

We have to show that $I \leq M$ where M is any finite positive constant.
First, integrating out with respect to $\boldsymbol{b}$ and using $\boldsymbol{w^T}(I - P_{\boldsymbol{X}_*})\boldsymbol{w} \geq 0$,

$$\begin{aligned} I_{\boldsymbol{b}} &= \int [\boldsymbol{\theta}|\boldsymbol{b},\boldsymbol{\gamma},\sigma_u^2,\boldsymbol{X}][\boldsymbol{b}]d\boldsymbol{b} \qquad \text{(A.2)} \\ &= (\sigma_u^2)^{-\frac{m}{2}} \int \exp\left[-\frac{1}{2\sigma_u^2}\sum_{i=1}^{m}\left(\theta_i - \boldsymbol{x_i^T b} - \boldsymbol{z_i^T}\gamma\right)^2\right]d\boldsymbol{b} \\ &= (\sigma_u^2)^{-\frac{m}{2}} \int \exp\left[-\frac{1}{2\sigma_u^2}\sum_{i=1}^{m}\left(w_i - \boldsymbol{x_i^T b}\right)^2\right]d\boldsymbol{b} \\ &= (\sigma_u^2)^{-\frac{m}{2}} \int \exp\left\{-\frac{1}{2\sigma_u^2}\boldsymbol{w^T}\left(I - P_{\boldsymbol{X}_*}\right)\boldsymbol{w}\right\}d\boldsymbol{b}(\sigma_u^2)^{\frac{2}{p}}|\boldsymbol{X_*^T X_*}|^{-\frac{1}{2}}(2\pi)^{\frac{m}{2}} \\ &\leq K_1 \cdot (\sigma_u^2)^{-\frac{(m-p)}{2}} \end{aligned}$$

where $P_{\boldsymbol{X}_*} = \boldsymbol{X}_*(\boldsymbol{X}_*{}^T\boldsymbol{X}_*)^{-}1\boldsymbol{X_*^T}, rank(\boldsymbol{X}_*) = p,$ *and* K_1 is constant.
Next, we consider integration with respect to σ_u^2,

$$\begin{aligned} I_{\sigma_u^2} &= \int (\sigma_u^2)^{-\frac{(m-p)}{2}}\left[\sigma_u^2\right]d\sigma_u^2 \qquad \text{(A.3)} \\ &= \int (\sigma_u^2)^{-\frac{(m-p)}{2}}(\sigma_u^2)^{-a_u+1}\exp\left(-b_u/\sigma_u^2\right)d\sigma_u^2 \\ &= \int (\sigma_u^2)^{-(a_u+\frac{m}{2}-\frac{p}{2})-1}\exp\left(-b_u/\sigma_u^2\right)d\sigma_u^2 \\ &= K_2 \end{aligned}$$

where K_2 is constant. Combining (A.2) and (A.3), we have

$$I \leq K_1K_2 \int \cdots \int \left[\boldsymbol{y}|\boldsymbol{\theta},\sigma_e^2\right]\left[\boldsymbol{X}|\sigma_\eta^2\right]\left[\gamma|\sigma_\gamma^2\right]\left[\sigma_e^2\right]\left[\sigma_\eta^2\right]\left[\sigma_\gamma^2\right]d\boldsymbol{\Omega^*} \qquad \text{(A.4)}$$

where $\boldsymbol{\Omega^*} = (\boldsymbol{\Omega}\boldsymbol{-b} - \sigma_u^2)$. Since all the components of the integrand in (A.4) have proper distribution, the above integral would be finite thus proving posterior propriety.

References

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

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties (with Discussion). *Statistical Science*, **11**, 89-121.

Ghosh, M. and Meeden, G. (1986). Empirical Bayes estimation in finite population sampling. *Journal of the American Statistical Association*, **81**, 1058-1062.

Ghosh, M. and Meeden, G. (1996). *Bayesian methods for finite population sampling,* Chapman & Hall, New York.

Ghosh, M. and Sinha, K. (2004). Empirical Bayes estimation in finite population sampling under functional measurement error models. *Journal of Statistical Planning & Inference*, **137**, 2759-2773.

Opsomer, J. D., Claeskens, G., Ranalli, M. G. and Breidt, F. J. (2008). Non-parametric small area estimation using penalized spline regression. *Journal of the Royal Statistical Society, Series B,* **70**, 265-286.

Rao, J. N. K. (2003). *Small area estimation*, John Wiley & Sons Inc, New York.