Robust small-area estimation under semi-parametric mixed models

Laura Dumitrescu [1] & J.N.K. Rao [2] & Sanjoy K. Sinha [3]

[1] *Department of Mathematics, Washington University in St. Louis, USA; E-mail: dumitrescu@math.wustl.edu*

[2] *School of Mathematics and Statistics, Carleton University, Canada; E-mail: jrao@math.carleton.ca*

[3] *School of Mathematics and Statistics, Carleton University, Canada; E-mail: sinha@math.carleton.ca*

Abstract. Model assisted estimators which make use of auxiliary information through linear models have been extensively used in small area estimation and their efficiency relies heavily on a correct specification of the linking model. Basic area-level and unit-level models have been considered and the classical approach is to use empirical best linear unbiased predictors (EBLUPs) of small area means. The EBLUPs are efficient under normality assumptions but their performance might be affected by the presence of representative outliers in the data, or departures from the normal distribution assumption of the random effects. In recent years several robust techniques have been developed for a unit- level linear mixed model. We relax the assumption of linear regression for the fixed part of the model and using a P-spline mixed model approximation we develop a robust empirical best linear unbiased predictor of the small area mean. To estimate its mean squared error, we adopt a conditional bootstrap approach. The performance of our method is demonstrated in a simulation study.

Keywords. Bootstrap, mean squared prediction error, outliers, small area mean, unit level model

1 Introduction

The interest in small area estimation has been increasing in the past decades and a popular model which is used is the nested error regression model. If y denotes the variable of interest, the model is of the form

$$y_{ij} = \mathbf{x}_{ij}^t \boldsymbol{\beta} + v_i + e_{ij}, \quad i = 1, \ldots, m; j = 1, \ldots, N_i, \tag{1}$$

where $\mathbf{x}_{ij} = (1, x_{1ij}, \ldots, x_{pij})^t$, m is the number of small areas, N_i is the number of population units j in area i, $v_i \sim_{iid} (0, \sigma_v^2)$ is the random small area effects and $e_{ij} \sim_{iid} (0, \sigma_e^2)$ are unit errors assumed to be independent of the v_i.

The population values of the auxiliary variables $x_1, \ldots, x_p$ for each area i are assumed to be known.

A sample of size n_i is collected from each area and the observed variables are y and the covariates $x_1, \ldots, x_p$. We assume that the sampling is non-informative so (1) also holds for the sample.

Given $\boldsymbol{\theta} = (\sigma_v^2, \sigma_e^2)^t$, the best linear unbiased predictor (BLUP) of the area mean $\bar{Y}_i$, which is also an empirical Bayes (EB) predictor under normality can be obtained from the sample model. If the true $\boldsymbol{\theta}$ is not known, the empirical BLUP (EBLUP) of $\bar{Y}_i$ is obtained by replacing $\boldsymbol{\theta}$ with a consistent estimator $\hat{\boldsymbol{\theta}}$.

The assumption of a parametric mean function of the form $m(\mathbf{x}, \boldsymbol{\beta}) = \mathbf{x}^t\boldsymbol{\beta}$ in (1) can be relaxed by assuming that the unknown function $m_0(x)$ can be well approximated by a penalized spline (P-spline) function with a truncated polynomial spline basis. Opsomer et al. (2008) extended the P-spline model to small area estimation by including the random area effects v_i and obtained the EBLUP of $\bar{Y}_i$

$$y_{ij} = \mathbf{x}_{ij}^t\boldsymbol{\beta} + \mathbf{w}_{ij}^t\mathbf{u} + v_i + e_{ij},\ j = 1, \ldots, n_i;\ i = 1, \ldots, m, \tag{2}$$

where $\mathbf{x}_{ij} = (1, x_{ij}, \ldots, x_{ij}^h)^t$, and $w_{ij} = \{(x_{ij} - q_1)_+^h, \ldots, (x_{ij} - q_K)_+^h\}^t = (w_{ij1}, \ldots, w_{ijK})^t$. They also studied the estimation of MSPE of the EBLUP.

The EBLUP can be sensitive to outliers in v_i and e_{ij}. In the case of EBLUP under the nested error regression model (1), Sinha and Rao (2009) developed a robust EBLUP method and demonstrated its advantages over the EBLUP in the presence of outliers in the random small area effects and/or unit level errors in the model. Our goal is to obtain a robust EBLUP (REBLUP) of the mean $\bar{Y}_i$ under a P-spline nested error regression model.

In matrix form, the P-spline model is a linear mixed model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{u} + \mathbf{Z}\mathbf{v} + \mathbf{e},$$

The algorithm for obtaining the REBLUPs of $\mathbf{u}$, $\mathbf{v}$ and robust estimators of $\boldsymbol{\beta}$, σ_u^2, σ_v^2 and σ_e^2 simultaneously is adapted from Fellner (1986).
Step 1. Given a starting value $\boldsymbol{\theta}^{(0)}$ of $\boldsymbol{\theta} = (\sigma_u^2, \sigma_v^2, \sigma_e^2)^t$, solve the mixed model equations to obtain $\boldsymbol{\beta}^{(0)}$, $\mathbf{u}^{(0)}$, $\mathbf{v}^{(0)}$ and compute $\mathbf{e}^{(0)} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(0)} - \mathbf{W}\mathbf{u}^{(0)} - \mathbf{Z}\mathbf{v}^{(0)}$.
Step 2. Using estimates from Step 1 update the estimate $\boldsymbol{\theta}^{(1)}$ of $\boldsymbol{\theta}$.
Step 3. Compute the pseudo-values

$$\begin{aligned} \tilde{\mathbf{y}} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{u} + \mathbf{Z}\mathbf{v} + \sigma_e\boldsymbol{\Psi}(\sigma_e^{-1}\mathbf{e}), \qquad (3) \\ \tilde{\mathbf{0}}_v &= \mathbf{v} - \sigma_v\boldsymbol{\Psi}(\sigma_v^{-1}\mathbf{v}), \qquad (4) \end{aligned}$$

where $\Psi(\mathbf{u}) = (\psi(u_1),\ \psi(u_2), \ldots)^t$, using $\boldsymbol{\beta}^{(0)}$, $\mathbf{u}^{(0)}$, $\mathbf{v}^{(0)}$ and $\mathbf{e}^{(0)}$ for $\boldsymbol{\beta}$, $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{e}$, and $\boldsymbol{\theta}^{(1)}$ for $\boldsymbol{\theta}$.

Step 4. Solve the robust mixed model equations

$$\begin{bmatrix} \sigma_e^{-2}\mathbf{X}^t\mathbf{X} & \sigma_e^{-2}\mathbf{X}^t\mathbf{W} & \sigma_e^{-2}\mathbf{X}^t\mathbf{Z} \\ \sigma_e^{-2}\mathbf{W}^t\mathbf{X} & \sigma_u^{-2}\mathbf{I}_K + \sigma_e^{-2}\mathbf{W}^t\mathbf{W} & \sigma_e^{-2}\mathbf{W}^t\mathbf{Z} \\ \sigma_e^{-2}\mathbf{Z}^t\mathbf{X} & \sigma_e^{-2}\mathbf{Z}^t\mathbf{W} & \sigma_v^{-2}\mathbf{I}_m + \sigma_e^{-2}\mathbf{Z}^t\mathbf{Z} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{u} \\ \mathbf{v} \end{bmatrix} = \begin{bmatrix} \sigma_e^{-2}\mathbf{X}^t\tilde{\mathbf{y}} \\ \sigma_e^{-2}\mathbf{W}^t\tilde{\mathbf{y}} \\ \sigma_e^{-2}\mathbf{Z}^t\tilde{\mathbf{y}} + \sigma_v^{-2}\tilde{\mathbf{0}}_v \end{bmatrix}, \tag{5}$$

using $\boldsymbol{\theta}^{(1)}$ and the corresponding pseudo-values obtained in Step 3. This leads to new values $\boldsymbol{\beta}^{(1)}$, $\mathbf{u}^{(1)}$, $\mathbf{v}^{(1)}$ and using these values compute $\mathbf{e}^{(1)} = \tilde{\mathbf{y}}^{(1)} - \mathbf{X}\boldsymbol{\beta}^{(1)} - \mathbf{W}\mathbf{u}^{(1)} - \mathbf{Z}\mathbf{v}^{(1)}$.

The steps are repeated until convergence is achieved. At convergence, robust estimators $\hat{\boldsymbol{\beta}}_F$, $\hat{\sigma}^2_{uF}$, $\hat{\sigma}^2_{vF}$ and $\hat{\sigma}^2_{eF}$ and REBLUPs $\hat{\mathbf{u}}_F$, $\hat{\mathbf{v}}_F$ are.

The equations (5) reduce to the mixed model equations when $\boldsymbol{\Psi}(\sigma_e^{-1}\mathbf{e}) = \sigma_e^{-1}\mathbf{e}$ and $\boldsymbol{\Psi}(\sigma_v^{-1}\mathbf{v}) = \sigma_v^{-1}\mathbf{v}$, and $\tilde{\mathbf{0}}_v$ reduces to $\mathbf{0}$.

The REBLUP of $\bar{Y}_i$ is given by

$$\hat{\mu}_{iF} = \frac{1}{N_i}\left(\sum_{j\in s_i} y_{ij} + \sum_{j\in \bar{s}_i} \hat{y}_{ijF}\right), \tag{6}$$

where $\hat{y}_{ijF} = \mathbf{x}^t_{ij}\hat{\boldsymbol{\beta}}_F + \mathbf{w}^t_{ij}\hat{\mathbf{u}}_F + \hat{v}_{iF}$.

For a non-sampled area l, the resulting predictor of $\bar{Y}_l$ is

$$\hat{\mu}_{lF} = \frac{1}{N_l}\sum_{j=1}^{N_l} \hat{y}_{ljF} = \bar{\mathbf{X}}^t_l\hat{\boldsymbol{\beta}}_F + \bar{\mathbf{W}}^t_l\hat{\mathbf{u}}_F. \tag{7}$$

$\hat{y}_{ljF} = \mathbf{x}^t_{lj}\hat{\boldsymbol{\beta}}_F + \mathbf{w}^t_{lj}\hat{\mathbf{u}}_F$, $j = 1, \ldots, N_l$ is the synthetic predictor of y_{lj}.

Due to the complex form of the REBLUP of the area mean $\bar{Y}_i$ and the lack of knowledge of the underlying distributions of $\mathbf{u}$, v_i and $\mathbf{e}_i$, MSPE estimation becomes difficult. We follow the Sinha-Rao method to obtain a conditional bootstrap estimator of MSPE of $\hat{\mu}_{iF}$ under the spline mixed model.

We generate $\mathbf{v}^*$ and $\mathbf{e}^*$ from $N(\mathbf{0}, \hat{\sigma}^2_{vF}\mathbf{I}_m)$ and $N(\mathbf{0}, \hat{\sigma}^2_{eF}\mathbf{I}_n)$ respectively, and obtain bootstrap responses $y^*_{ij} = \mathbf{x}^t_{ij}\hat{\boldsymbol{\beta}}_F + \mathbf{w}^t_{ij}\hat{\mathbf{u}}_F + v^*_i + e^*_{ij}$, $j = 1, \ldots, N_i$ and $i = 1, \ldots, m$ that are free of outliers. Using the corresponding bootstrap sample data $\{(y^*_{ij}, \mathbf{x}_{ij}, \mathbf{w}_{ij}, j \in s_i; i = 1, \ldots, m)\}$ we obtain bootstrap estimates $\hat{\boldsymbol{\beta}}^*$, $\hat{\mathbf{u}}^*$ and $\hat{v}^*_i$ and the predicted values $\hat{y}^*_{ij} = \mathbf{x}^t_{ij}\hat{\boldsymbol{\beta}}^* + \mathbf{w}^t_{ij}\hat{\mathbf{u}}^* + \hat{v}^*_i$ for $j \in \bar{s}_i$, where $\hat{\boldsymbol{\beta}}^*$ is the BLUE of $\boldsymbol{\beta}$ and $\hat{\mathbf{u}}^*$ and $\hat{v}^*_i$ are the EBLUPs of $\mathbf{u}$ and v_i, respectively. The resulting bootstrap REBLUP of $\bar{Y}_i$ is

$$\hat{\mu}^*_i = \frac{1}{N_i}\left(\sum_{j\in s_i} y^*_{ij} + \sum_{j\in \bar{s}_i} \hat{y}^*_{ij}\right)$$

and the bootstrap population mean is $\bar{Y}^*_i = N_i^{-1}\sum_{j=1}^{N_i} y^*_{ij}$. The use of EBLUPs rather than REBLUPs is reasonable because the bootstrap sample is free of outliers. Repeating the bootstrap operation B times, the bootstrap estimator of the MSPE of $\hat{\mu}_{iF}$ is given by

$$\hat{M}_{\text{boot}}(\hat{\mu}_{iF}) = \frac{1}{B}\sum_{i=1}^{B}\{\hat{\mu}_i^*(b) - \bar{Y}_i^*(b)\}^2, \tag{8}$$

where $\hat{\mu}_i^*(b)$ and $\bar{Y}_i^*(b)$ are the values of $\hat{\mu}_i^*$ and $\bar{Y}_i^*$ for the b-th bootstrap replicate and $\hat{\mu}_{iF}$ is given by (6).

We report some results of a simulation study on the performance of the proposed P-spline REBLUP of a small area mean. We generate samples according to a true semi-parametric nested error model with a linear, quadratic, and exponential (1) mean functions and look at the performance of the proposed methodunder diffeent contamination schemes.

Bibliography

[1] Fellner, W. H. (1986). Robust estimation of variance components. *Technometrics*, 28, 51–60.

[2] Opsomer, J. P., Claeskens, G., Ranalli, M. G., Kauemann, 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.

[3] Rao, J.N.K, Sinha, S. K., Dumitrescu, L. (2014) Robust small area estimation under semi-parametric mixed models, *The Canadian Journal of Statistics*, 42(1), 126-141.

[4] Sinha, S. K. and Rao, J. N. K. (2009). Robust small area estimation. *The Canadian Journal of Statistics*, 37, 381–399.