An Outlier Robust Block Bootstrap for Small Area Estimation

Payam Mokhtarian and Ray Chambers

National Institute for Applied Statistics Research Australia

University of Wollongong

The First Asian ISI Satellite Meeting on Small Area Estimation

1 – 4 September 2013, Chulalongkorn University, Bangkok, Thailand

NIASRA
NATIONAL INSTITUTE FOR APPLIED STATISTICS RESEARCH AUSTRALIA

UNIVERSITY OF WOLLONGONG

Overview

- Introduction, Background and Motivation

Part I

- Assumptions and Model Specification
- Outlier Robust LMM Fitting
- The Outlier Robust Block Bootstrap

Part II

- Robust Small Area Estimation
- MSE Estimation
- Numerical Results
- Concluding Remarks

Introduction and Background

- Outliers in data are a well-known problem when fitting models
- Estimates of the model parameters and predictions of population quantities become unstable in the presence of outliers
- Accurate estimation of variance components is a challenge when there are outliers in the sample data
- In order to tackle this issue, parameter estimating functions are usually modified to make them less outlier sensitive (M-estimation)
  - Richardson and Welsh (1995): Robust REML for mixed linear models

- We propose an outlier robust Monte Carlo (bounded bootstrap) method to deal with the influence of outliers on estimates of mixed model parameters (Chambers and Chandra, 2013)
- Method leads to more reliable mixed model parameter estimates than comparable outlier robust approaches proposed in the literature
- This approach is not hard to implement since it based on bootstrapping
- We provide a Theorem on the asymptotic bias of the proposed approach

- Natural extension of this Robust Random Effect Block (RREB) bootstrap approach is to Small Area Estimation
- Three different outlier robust predictors of a small area mean are proposed
- Two types of Mean Squared Error (MSE) estimator for the proposed REBB-based predictors are proposed
- Numerical results indicate that the proposed robust method is stable and leads to a reliable small area mean predictor with a smaller MSE

PART I

An Outlier Robust Method for Estimating the Parameters of a Linear Mixed Model

Assumptions and Model Specification

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + u_i + e_{ij}\,, i = 1,\ldots,D\,,\ j = 1,\ldots,N_i$$

- $\mathbf{y}_i = (y_{i1},\ldots,y_{iN_i})^T$ is $N_i \times 1$ vector of variable of interest
- $\mathbf{X}_i = [\mathbf{x}_{i1}\ldots\mathbf{x}_{iN_i}]^T$ is $N_i \times p$ covariate matrix
- $\mathbf{u}$ is the vector of area effects (level 2) and $\mathbf{e}_i$ is $N_i \times 1$ vector of individual effects (level 1)
- $\mathbf{u} \sim N(\mathbf{0}, \sigma_u^2 \mathbf{I}_D)$ and $\mathbf{e}_i \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_{N_i})$, where $\mathbf{u} \perp \mathbf{e}_i$
- Fixed effects: $\boldsymbol{\beta}$; variance components: $\boldsymbol{\theta} = (\sigma_u^2, \sigma_e^2)^T$
- Covariance matrix of $\mathbf{y}_i$: $\mathbf{V}_i = \mathbf{V}_i(\boldsymbol{\theta}) = \sigma_e^2 \mathbf{I}_{N_i} + \mathbf{1}_{N_i} \mathbf{1}_{N_i}^T \sigma_u^2$

Outlier Robust REML Estimation Equations

- Richardson & Welsh (1995): Bounded estimating functions

$$\sum_{i=1}^{D} \mathbf{X}_i^T \mathbf{V}_i^{-1/2} \psi(\mathbf{r}_i(\boldsymbol{\theta})) = \mathbf{0} \qquad \text{(A)}$$

$$\sum_{i=1}^{D} \left\{ \psi^T(\mathbf{r}_i(\boldsymbol{\theta})) \mathbf{V}_i^{-1} \frac{\partial \mathbf{V}_i}{\partial \theta_l} \mathbf{V}_i^{-1} \psi(\mathbf{r}_i(\boldsymbol{\theta})) - tr\left(\mathbf{K}_{2i} \mathbf{P}_i \frac{\partial \mathbf{V}_i}{\partial \theta_l} \right) \right\} = \mathbf{0} \qquad \text{(B)}$$

- Iterative methods used to solve the estimating equations become numerically unstable as the number of variance components increases
- Estimation of 'non-outlier' variance components is biased when outliers are present - although this bias is less than that of REML

Bootstrap Model Fitting

- Chambers and Chandra (2013) developed a procedure to fit a linear mixed model using a random effect block bootstrap (REB)
  - REB is robust to failure of the level 1 independence assumptions of the mixed model
- We propose an outlier robust extension of the REB idea that can be used to fit a linear mixed model in the presence of both level 2 and level 1 outliers

based on bounding the influence of outliers on the bootstrap distributions of the marginal residuals

Outlier Robust Block Bootstrap (RREB)

Given the hierarchical structure of the linear mixed model we can calculate

1. Marginal residuals:

$$r_{ij} = y_{ij} - \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}};\ (\mathbf{r} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

- Group average residuals:

$$\overline{r}_{i.} = n_i^{-1} \sum_{j=1}^{n_i} r_{ij}\ ;\ \overline{\mathbf{r}}^{(2)} = \left\{ \overline{r}_{i.} \right\}$$

- Standardised group average residuals:

$$\mathbf{r}^{(2)-\mathrm{C}} = \overline{\mathbf{r}}^{(2)} - av(\overline{\mathbf{r}}^{(2)})\mathbf{1}_D$$

$$\mathbf{r}^{(2)-\mathrm{SC}} = \left\{ D^{-1} (\mathbf{r}^{(2)-\mathrm{C}})^T \mathbf{r}^{(2)-\mathrm{C}} \right\}^{-1/2} \mathbf{r}^{(2)-\mathrm{C}} \hat{\sigma}_u$$

- Outlier robust group level (level 2) residuals:

$$\mathbf{r}^{(2)\mathrm{R}} = \psi_2\left(\mathbf{r}^{(2)-\mathrm{SC}}\right);\ c_2 = 2\hat{\sigma}_u$$

- Standardised individual level residuals:

$$\mathbf{r}^{(1)-\mathrm{C}} = (\mathbf{r} - \mathbf{r}^{(2)\mathrm{R}} \otimes \mathbf{1}_{n_i}) - av(\mathbf{r} - \mathbf{r}^{(2)\mathrm{R}} \otimes \mathbf{1}_{n_i})$$

$$\mathbf{r}^{(1)-\mathrm{SC}} = \left\{ n^{-1} (\mathbf{r}^{(1)-\mathrm{C}})^T \mathbf{r}^{(1)-\mathrm{C}} \right\}^{-1/2} \mathbf{r}^{(1)-\mathrm{C}} \hat{\sigma}_e$$

- Outlier robust individual level (level 1) residuals:

$$\mathbf{r}^{(1)\mathrm{R}} = (\mathbf{r}_i^{(1)\mathrm{R}}) = \psi_1\left(\mathbf{r}^{(1)-\mathrm{SC}}\right);\ c_1 = 2\hat{\sigma}_e$$

2. Bootstrap errors defined by sampling with replacement from each set of robust residuals (independently at level 2, block sampling at level 1)

$$\mathbf{r}^{*(2)\mathrm{R}} = \left(r_i^{*(2)\mathrm{R}} \right) = srswr\left(\mathbf{r}^{(2)\mathrm{R}}, D \right)$$

$$\mathbf{r}_i^{*(1)R} = \left(r_{ij}^{*(1)R} \right) = srswr\left(\mathbf{r}^{(1)R}_{j=srswr\left(\{1,\square,D\},1\right)}, n_i \right)$$

$$\mathbf{r}^{*(1)\mathrm{R}} = \left(\mathbf{r}_i^{*(1)\mathrm{R}} \right)$$

3. Robust bootstrap sample data $(y_{ij}^{*\mathrm{R}}, \mathbf{x}_{ij})$ are generated via

$$y_{ij}^{*\mathrm{R}} = \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}} + r_i^{*(2)\mathrm{R}} + r_{ij}^{*(1)\mathrm{R}}$$

4. A two-level linear mixed model is fitted to these bootstrap sample data to obtain bootstrap parameter estimates

$$\hat{\boldsymbol{\phi}}^{*\mathrm{R}} = \left(\hat{\boldsymbol{\beta}}^{*\mathrm{R}}, \hat{\sigma}_u^{2*\mathrm{R}}, \hat{\sigma}_e^{2*\mathrm{R}} \right)$$

5. Repeat to obtain *B* sets of bootstrap parameter estimates

- Marginal residuals used in the bootstrapping process assume $\hat{\beta}$ is consistent estimator for fixed effects (here REML or RREML)
- The variance component estimates for the both level 1 and level 2 effects can be either REML type or RREML type
- In the contaminated case (the case of most interest) using RREML estimates for the estimated variance components used in the standardisation step leads to less biased RREB variance components estimates
- Note that RREB variance components estimates are still significantly biased - but this bias is much smaller than that of RREML
- An adaptive algorithm is proposed which reduces this bias of the RREB variance components estimates

An Adaptive Robust Block Bootstrap (ARREB)

- Iterate the RREB bootstrap using, $\hat{\boldsymbol{\phi}}^{\text{RREB}} = \left(\hat{\boldsymbol{\beta}}^{\text{RREB}}, \hat{\sigma}_u^{2\text{RREB}}, \hat{\sigma}_e^{2\text{RREB}}\right)$ from previous iteration as input to current iteration
  - $\hat{\boldsymbol{\beta}}^{\text{RREB}}$ replaces $\hat{\boldsymbol{\beta}}$ when calculating new marginal residuals
  - $\hat{\sigma}_u^{2\text{RREB}}$ and $\hat{\sigma}_e^{2\text{RREB}}$ replace $\hat{\sigma}_u^2$ and $\hat{\sigma}_e^2$ when re-scaling the level 2 and level 1 residuals
  - Subsequence steps in the RREB algorithm are unchanged
- Iterations continue until $\left\|\hat{\boldsymbol{\phi}}^{\text{ARREB}-(s)} - \hat{\boldsymbol{\phi}}^{\text{ARREB}-(s-1)}\right\| \leq \delta$. In our numerical evaluations we set $\delta = 10^{-3}$

PART II

Using RREB for Outlier Robust Small Area Estimation

Outlier Robust Small Area Estimation

- Area-specific sample sizes are small and so outliers in the sample data have a significant effect on inference for any particular area
- Chambers and Tzavidis (2006) proposed an *M*-quantile approach that is robust to the presence of individual (level 1) outliers
- Sinha and Rao (2009) proposed an outlier robust EBLUP (REBLUP) using the robust model fitting approach of Richardson and Welsh (1995), as well as a bootstrap MSE estimator (BOOT)
- Chambers et al (2013) proposed a bias-corrected version of both the REBLUP and the *M*-quantile estimators. They also provided two analytical MSE estimators (CCT, CCST) for these robust SAE methods

- Under the assumed linear mixed model, the EBLUP of the area *i* mean $\bar{y}_i$ is:

$$\hat{\bar{y}}_i^{\text{EBLUP}} = N_i^{-1}\left\{n_i\bar{y}_{si} + (N_i - n_i)\hat{\bar{y}}_{ri}\right\};\ \hat{\bar{y}}_{ri} = \bar{\mathbf{x}}_{ri}^T\hat{\beta} + \hat{u}_i$$

- REBLUP of the area *i* mean $\bar{y}_i$ proposed by Sinha and Rao (2009) is:

$$\hat{\bar{y}}_i^{\text{REBLUP}} = N_i^{-1}\left\{n_i\bar{y}_{si} + (N_i - n_i)\hat{\bar{y}}_{ri}^{\text{SR}}\right\};\ \hat{\bar{y}}_{ri}^{\text{SR}} = \bar{\mathbf{x}}_{ri}^T\hat{\beta}^{\text{SR}} + \hat{u}_i^{\text{SR}}$$

  where the unknown parameters are estimated using the robust approach proposed by Richardson and Welsh (1995)

- Algorithms used to calculate the REBLUP are unstable. Also MSE estimates are not reliable
- We use the RREB approach to obtain more accurate and easily implemented small area mean estimates and associated MSE estimates

RREB-based Small Area Estimation

- RREB-based EBLUP of the area *i* mean $\overline{y}_i$ is:

$$\hat{\overline{y}}_i^{\text{RREB}} = N_i^{-1}\left\{n_i\overline{y}_{si} + (N_i - n_i)\hat{\overline{y}}_{ri}^{\text{RREB}}\right\};\ \hat{\overline{y}}_{ri}^{\text{RREB}} = \overline{\mathbf{x}}_{ri}^T\hat{\beta}^{\text{RREB}} + \hat{u}_i^{\text{RREB}}$$

- We investigate three version of $\hat{u}_i^{\text{RREB}}$ depending on the type of bootstrap averaging used to obtain this predicted value

$$\hat{u}_i^{\text{RREB-1}} = B^{-1}\sum_{b=1}^{B}\left\{\left(n_i^{-1}\hat{\sigma}_e^{2(b)\text{RREB}} + \hat{\sigma}_u^{2(b)\text{RREB}}\right)^{-1}\hat{\sigma}_u^{2(b)\text{RREB}}\left(\overline{y}_{si} - \overline{\mathbf{x}}_{si}^T\hat{\beta}^{(b)\text{RREB}}\right)\right\}$$

$$\hat{u}_i^{\text{RREB-2}} = \left\{B^{-1}\sum_{b=1}^{B}\left(n_i^{-1}\hat{\sigma}_e^{2(b)\text{RREB}} + \hat{\sigma}_u^{2(b)\text{RREB}}\right)^{-1}\hat{\sigma}_u^{2(b)\text{RREB}}\right\}\left\{\overline{y}_{si} - \overline{\mathbf{x}}_{si}^T\hat{\beta}^{\text{RREB}}\right\}$$

$$\hat{u}_i^{\text{RREB-3}} = \left\{\left(n_i^{-1}\hat{\sigma}_e^{2\text{RREB}} + \hat{\sigma}_u^{2\text{RREB}}\right)^{-1}\hat{\sigma}_u^{2\text{RREB}}\right\}\left\{\overline{y}_{si} - \overline{\mathbf{x}}_{si}^T\hat{\beta}^{\text{RREB}}\right\}$$

- We compare these alternatives in our numerical evaluations

RREB-based MSE Estimation

- We propose two approaches to estimating the MSE of the RREB-based predictor of the small area mean
  - Using the Prasad and Rao (1990) method of MSE estimation
  - Using the observed variability in the RREB bootstrap replications

Plug-in Prasad-Rao type MSE estimator (PR-I)

$$\mathrm{MSE}^{\mathrm{PR}} = g_{1i}(\hat{\boldsymbol{\theta}}^{\mathrm{REML}}) + g_{2i}(\hat{\boldsymbol{\theta}}^{\mathrm{REML}}) + 2g_{3i}(\hat{\boldsymbol{\phi}}^{\mathrm{REML}})$$

where each component depends on the REML estimates of the variance components and their estimated variances and covariance, with

$$\hat{\boldsymbol{\phi}}^{\mathrm{REML}} = \left(\hat{\boldsymbol{\theta}}^{\mathrm{REML}}, \mathrm{v}_u(\hat{\boldsymbol{\theta}}^{\mathrm{REML}}), \mathrm{v}_e(\hat{\boldsymbol{\theta}}^{\mathrm{REML}}), \mathrm{c}_{ue}(\hat{\boldsymbol{\theta}}^{\mathrm{REML}})\right)$$

- g_1 and g_2 depend only on $\hat{\boldsymbol{\theta}}^{\mathrm{REML}}$, but g_3 depends on $\hat{\boldsymbol{\phi}}^{\mathrm{REML}}$
- Plug-in RREB version of PR MSE estimator uses

$$\hat{\boldsymbol{\phi}}^{\text{RREB-I}} = \left(\hat{\boldsymbol{\theta}}^{\mathrm{RREB}}, \mathrm{v}_u(\hat{\boldsymbol{\theta}}^{\mathrm{RREB}}), \mathrm{v}_e(\hat{\boldsymbol{\theta}}^{\mathrm{RREB}}), \mathrm{c}_{ue}(\hat{\boldsymbol{\theta}}^{\mathrm{RREB}})\right)$$

$$\mathrm{MSE}^{\mathrm{PR-I}}(\hat{\bar{y}}_i^{\mathrm{RREB}}) = g_{1i}(\hat{\boldsymbol{\theta}}^{\mathrm{RREB}}) + g_{2i}(\hat{\boldsymbol{\theta}}^{\mathrm{RREB}}) + 2g_{3i}(\hat{\boldsymbol{\phi}}^{\text{RREB-I}})$$

Bootstrap-based Prasad-Rao type MSE estimator (PR-II)

- We use bootstrap estimates of the variances and covariance of the RREB estimates of the variance components

$$\mathrm{v}_u^{\mathrm{RREB}} = B^{-1}\sum_{b=1}^{B}\left(\hat{\sigma}_u^{2(b)\mathrm{RREB}} - \hat{\sigma}_u^{2\mathrm{RREB}}\right)^2$$

$$\mathrm{v}_e^{\mathrm{RREB}} = B^{-1}\sum_{b=1}^{B}\left(\hat{\sigma}_e^{2(b)\mathrm{RREB}} - \hat{\sigma}_e^{2\mathrm{RREB}}\right)^2$$

$$c_{ue}^{\mathrm{RREB}} = B^{-1}\sum_{b=1}^{B}\left(\hat{\sigma}_u^{2(b)\mathrm{RREB}} - \hat{\sigma}_u^{2\mathrm{RREB}}\right)\left(\hat{\sigma}_e^{2(b)\mathrm{RREB}} - \hat{\sigma}_e^{2\mathrm{RREB}}\right)$$

leading to $\hat{\boldsymbol{\phi}}^{\text{RREB-II}} = \left(\hat{\boldsymbol{\theta}}^{\mathrm{RREB}}, \mathrm{v}_u^{\mathrm{RREB}}, \mathrm{v}_e^{\mathrm{RREB}}, \mathrm{c}_{ue}^{\mathrm{RREB}}\right)$

and

$$\mathrm{MSE}^{\mathrm{PR-II}}(\hat{\bar{y}}_i^{\mathrm{RREB}}) = g_{1i}(\hat{\boldsymbol{\theta}}^{\mathrm{RREB}}) + g_{2i}(\hat{\boldsymbol{\theta}}^{\mathrm{RREB}}) + 2g_{3i}(\hat{\boldsymbol{\phi}}^{\text{RREB-II}})$$

Bootstrap MSE estimator (RREB)

- This MSE estimator uses the observed bootstrap variability of $\hat{\bar{y}}_i^{\mathrm{RREB}}$, and is given by

$$\mathrm{MSE}^{\mathrm{RREB}}(\hat{\bar{y}}_i^{\mathrm{RREB}}) = B^{-1}\sum_{b=1}^{B}\left(\hat{\bar{y}}_i^{(b)\mathrm{RREB}} - \hat{\bar{y}}_i^{\mathrm{RREB}}\right)^2$$

Model Based Simulation

Same model and simulation set-up as in Part I model based simulation

Model-based simulation results for predictors of small area means

Predictor	*Results (%) for the scenarios and areas*					
	[0,0] 1-40	[0,**e**] 1-40	[**u**,0] 1-36	[**u**,0] 37-40	[**u**,**e**] 1-36	[**u**,**e**] 37-40
			Median values of RB			
EBLUP	0.02	-0.20	0.10	-0.54	0.17	-1.59
REBLUP	0.03	-0.39	0.11	-0.47	-0.30	-1.00
RREB-1	0.04	-0.33	0.91	-6.71	0.63	-6.81
RREB-2	0.02	-0.18	0.08	-0.42	0.10	-0.78
RREB-3	0.04	0.32	0.85	-6.70	0.58	-6.78
			Median values of RRMSE			
EBLUP	0.81	1.22	0.85	0.97	1.37	2.36
REBLUP	0.82	1.01	0.84	1.02	0.99	1.44
RREB-1	1.71	1.77	1.92	7.55	1.84	7.61
RREB-2	0.81	1.19	0.85	0.97	1.02	1.42
RREB-3	0.83	1.25	0.82	2.18	1.39	2.21

Model-based simulation results for relative bias of RMSE estimators

Predictor	*MSE Estimator*	*Median values of RB Results (%) for the scenarios and areas*					
		[0,0] 1-40	[0,**e**] 1-40	[**u**,0] 1-36	[**u**,0] 37-40	[**u**,**e**] 1-36	[**u**,**e**] 37-40
EBLUP	PR	-0.34	1.74	3.82	-17.31	11.32	-40.86
	CCT	3.61	31.24	1.55	2.15	5.95	-3.05
	CCST	0.55	31.22	-3.91	-0.30	2.96	-4.17
REBLUP	CCT	-17.71	-1576	-20.24	-34.79	-19.51	-36.63
	CCST	-2.01	-8.46	-3.58	-3.58	-7.91	-22.51
	BOOT	-1.19	-4.42	-19.42	-19.42	11.37	-31.44
RREB-2	PR-I	0.71	0.81	2.14	2.38	3.01	3.20
	PR-II	0.42	-0.65	2.02	2.18	2.58	2.69
	RREB	-0.91	-0.89	-0.94	-0.82	-0.95	-0.88

Model-based simulation results for performance of RMSE estimators

Predictor	*MSE Estimator*	*Median values of RRMSE Results (%) for the scenarios and areas*					
		[0,0] 1-40	[0,**e**] 1-40	[**u**,0] 1-36	[**u**,0] 37-40	[**u**,**e**] 1-36	[**u**,**e**] 37-40
EBLUP	PR	6.24	18.57	7.20	17.90	22.28	43.19
	CCT	31.51	76.20	31.25	28.37	61.57	51.30
	CCST	22.92	66.27	7.68	18.98	27.15	39.13
REBLUP	CCT	29.52	30.82	28.67	28.58	29.00	38.70
	CCST	27.86	28.47	20.89	22.87	20.25	29.24
	BOOT	10.27	34.92	10.67	14.62	16.61	33.04
RREB-2	PR-I	7.35	20.11	8.18	16.64	18.08	23.66
	PR-II	7.81	21.97	8.87	16.71	18.67	23.82
	RREB	5.39	10.46	5.59	9.88	10.11	12.84

Current & Future Research

- Currently investigating application of RREB to clustered data with spatial similarity with the aim of obtaining more efficient estimates of variance components and spatial correlation parameters in the presence of outliers
  - Timo Schmid, Freie University Berlin
- Application of this Spatial RREB to Small Area Estimation where small area have spatial area structure is a potential application
- Bias-corrected version of RREB needs to be developed (Chambers *et al* 2013)
- Extending the RREB idea to fitting generalized linear mixed models for count and binary data is a topic for further research

Main References

[1] Chambers, R, Chandra, H., Salvati, N. and Tzavidis, N. (2013). **Outlier Robust Small Area Estimation**, *Journal of the Royal Statistical Society, Series B*. **75**, part 5. 1-23.

[2] Chambers, R. and Chandra, H. (2013). **Random Effect Block Bootstrap for Clustered Data**, *Journal of Computational and Graphical Statistics*, **22**, 452-470.

[3] Chambers, R. and Tzavidis, N. (2006). ***M*-quantile Models for Small Area Estimation**, *Biometrika*, 93, 255-268.

[4] Sinha, S.K. and Rao, J.N.K. (2009). **Robust Small Area Estimation**, *Canadian Journal of Statistics*, **37**, 381-399.

[5] Richardson, A.M. and Welsh, A.H. (1995). **Robust Restricted Maximum Likelihood in Mixed Linear Models**, *Biometrics*, **52**, 1429-1439.