An Application of Parametric Bootstrap Method in Small Area Estimation Problem

Huilin Li
University of Maryland, College Park

Abstract

In this paper, we apply the recently developed parametric bootstrap method in constructing prediction intervals of small area means for the well-known small area models: the Fay-Herriot model. Using a Monte Carlo simulation study, we compare our method with different rival methods in terms of coverage probabilities and average lengths. We then demonstrate the utility of the parametric bootstrap method by analyzing a real life dataset.

Keywords: Parametric bootstrap, Prediction intervals, Fay-Herriot model, ADM estimator.

1 Introduction

The term small domain or area typically refers to a population for which reliable statistics of interest cannot be produced due to certain limitations of the available data. Small-area statistics are needed in regional planning, apportioning congressional seats, and fund allocation in many government programs and thus the importance of producing reliable small-area statistics cannot be overemphasized.

In a sample survey, the sampling design usually aims to provide reliable data for large areas and leads very few samples or even no sample in the certain small areas. Thus the design-based small area estimators are unreliable, and various model-based indirect methods are considered to combine information from the related sources to derive reliable small area estimators. Mixed effects models are widely used in small area estimation. Such a model includes certain fixed effects to explain the between area variations in the interested variables, and area specific random effects to account for the between area variations not explained by the fixed effects. The Fay-Herriot Model, a mixed regression model is a well-known mixed models in the small area estimation.

In this paper, we concern the interval prediction under a mixed regression model in the context of small area estimation. For the i^{th} small area mean θ_i, we are interested in the $100(1-\alpha)\%$ prediction interval $PI_i(\mathbf{Y})$, which satisfies $Pr(\theta_i \in PI_i(\mathbf{Y})) = 1-\alpha$, where the probability is with respect to the marginal distribution of $\mathbf{Y}$.

In Section 2, we introduce the parametric bootstrap prediction interval proposed by Chatterjee et al.(2006) for different small area parameters of interest using the general linear mixed model. In Section 3, we present the ADM (adjustment for density maximization) method for making inferences about the random effect parameters. This new estimators can improve the prediction interval constructed by parametric bootstrap method. We compare the different prediction intervals under the simple Fay-Herriot model in Section 4. In Section 5, we present results from a Monte Carlo simulation study.

2 Parametric Bootstrap Prediction Interval

The following general linear mixed model covers a wide range of multi-level small area models:

$$Y = X\beta + Zv + e, \tag{1}$$

where $X(n \times p)$ and $Z(n \times q)$ are known matrices, $Y(n \times 1)$ is the observed data, v and e are independently distributed with $N(0, D)$ and $N(0, R)$ separately. $D = D(\psi)(q \times q)$ and $R = R(\psi)(n \times n)$ depend on $\psi = (\psi_0, \psi_1, \cdots, \psi_k)^{'}$, a $(k+1) \times 1$ vector of fixed variance components. Note that the dispersion matrix of the observed data Y is given by $\Sigma(\psi) = R + ZDZ^{'}$.

We are interested in investigating the distribution of $T = c^T(X\beta + Zv)$, where c is any fixed and known $(n \times 1)$ vector. When $c^T = (0, 0, \cdots, 1, \cdots, 0)$ where only the i^{th} element is 1, T represents the i^{th} small area mean. When $\phi = (\beta, \psi)$ are known,

$$T|Y \sim N(\mu_T, \sigma_T^2).$$

where μ_T and σ_T^2 are the posterior mean and variance of T given Y respectively,

$$\mu_T = c^{'}X\beta + c^{'}ZDZ^{'}\Sigma^{-1}(Y - X\beta) \tag{2}$$

$$= c^{'}R\Sigma^{-1}X\beta + c^{'}ZDZ^{'}\Sigma^{-1}Y, \tag{3}$$

$$\sigma_T^2 = c^{'}Z(D - DZ^{'}\Sigma^{-1}ZD)Z^{'}c.$$

Naturally we can construct the prediction interval for T as:

$$PI(t) = [\mu_T - z_{\alpha/2}\sigma_T, \mu_T + z_{\alpha/2}\sigma_T],$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ percent point of the normal distribution.

In practice ϕ are usually unknown and need to be estimated from the marginal distribution of Y. Then an EBLUP of T is $\hat{\mu}_T$, obtained from μ_T with ϕ replaced by $\hat{\phi}$. Also a naive variance estimator of T is given by $\hat{\sigma}_T^2$, obtained by replacing ϕ with $\hat{\phi}$. A naive prediction interval for T is constructed as:

$$PI_{naive}(t) = [\hat{\mu}_T - z_{\alpha/2}\hat{\sigma}_T, \hat{\mu}_T + z_{\alpha/2}\hat{\sigma}_T].$$

Obviously, this prediction interval is usually too narrow to attain the target converge probability due to the lack of the variability caused by estimation of ϕ in $\hat{\sigma}_T$.

Chatterjee et al.(2006) proposed a parametric bootstrap prediction interval. In their method, they employed the $\hat{\mu}_T$ and $\hat{\sigma}_T^2$ to construct prediction interval. Since $(T-\hat{\mu}_T)/\hat{\sigma}_T$ is no longer a normal distribution, in which case $z_{\alpha/2}$ is not a proper cut-off point, they find the cut-off points t from the parametric bootstrap samples. Their prediction interval is given as:

$$PI_{boot}(t) = [\hat{\mu}_T - t_1\hat{\sigma}_T, \hat{\mu}_T + t_2\hat{\sigma}_T], t \in R$$

In the above, $(t_1, t_2$ is obtained using

$$P^*\{T^* \in [\hat{\mu}_T^* - t_1\hat{\sigma}_T^*, \hat{\mu}_T^* + t_2\hat{\sigma}_T^*]\} = 1-\alpha$$

where the probability P^* is with respect to the parametric bootstrap distribution and T^*, $\hat{\mu}_T^*$ and $\hat{\sigma}_T^*$ are like T, $\hat{\mu}_T$ and $\hat{\sigma}_T$, except that bootstrap samples are used in place of original sample. The coverage probability of this prediction interval is accurate up to $O(m^{-3/2})$.

3 ADM method

Motivated by the Morris' discussion of Jiang and Lahiri (2006), we consider the application of ADM(adjustment for density maximization, see Morris(1988)) method in the interval prediction. ADM method is designed to approximate the posterior means and variances corresponding to a superharmonic prior on the between group variance component. In this section we describe how to use ADM method to estimate the between area variance R instead using the ML/REML methods and apply it to the previous parametric bootstrap interval prediction.

In the practice, the variance components are usually unknown. MLE and REML methods are widely used to estimate them. However, when the number of small area is limited, MLE works poorly. Morris suggested the ADM method to estimate shrinkage factors. Regard the REML adjusted likelihood $L(R)$ as the marginal posterior density of R, one can get the posterior mode by maximizing $L(R)$. $L(R)$ is right-skewed, noticeably so if m is not large, so the mean of R exceeds its mode. The ADM method is to maximize $R*L(R)$ instead $L(R)$, which gives better approximation than does the mode. This multiplication by R corrects for underestimation of R and also for convexity of B_i. Unlike estimates from MLE and REML maximizations, ADM estimates of R are always positive and shrinkage factors $B = D/(D+R)$ automatically are constrained to $[0, 1]$.

Tang(2002) shows that the point and interval estimates stemming from an ADM application with the superharmonic prior have much better accuracies and coverages in frequency evaluations than MLE and REML. In this paper, we will compare the parametric bootstrap prediction interval prediction using the REML estimator and the ADM estimator.

4 The Fey-Herriot Model

Fay and Herriot (1979) used the following two level model to estimate per capita income (PCI) for small places in the United States with population less than 1,000.

Level 1: $y_i|\theta_i \overset{ind}{\sim} N[\theta_i, D_i], i = 1, \cdots, m,;$

Level 2: $\theta_i \overset{iid}{\sim} N[x_i^{'}\beta, A], i = 1, \cdots, m,.$

The above two level model can be written as the following mixed model:

$$y_i = \theta_i + e_i = x_i^{'}\beta + v_i + e_i, i = 1, \cdots, m,$$

where $v_i \sim N(0, A)$ and $e_i \overset{ind}{\sim} N(0, D_i)$. The area-specific random effect $v_i \overset{ind}{\sim} N(0, A)$ is used to link the true small area means θ_i to a vector of p known auxiliary variables x_i, which are often obtained from various administrative and census records. The paremeters β and A are generally unknown and are estimated from the marginal distribution of y. The sampling variances D_i are usually assumed known.

We are interested in obtaining the prediction interval for the true small area means $\theta_i = x_i^T\beta + v_i$. We consider the following 4 methods.

Method 1: Direct Method

The method is based on the data [Level 1] only and does not use any prior model [Level 2] information. The direct prediction interval θ_i is given by

$$PI_i^D(\alpha) = [y_i - z_{\alpha/2}\sqrt{D_i}, y_i + z_{\alpha/2}\sqrt{D_i}]$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ percent point of $N(0,1)$. Obviously, for this prediction interval, the coverage probability is $1-\alpha$. However, it is not efficient since its average length is too large to make any reasonable conclusion. This is due to the high variability of the point predictor y_i.

Method 2: Synthetic Parametric Bootstrap Method

When β and A are known, we can construct the prediction interval of θ_i as $[x_i^{'}\beta \pm z_{\alpha/2}\sqrt{A}]$ without using the data. It relies totally on the prior model and hence is synthetic.

When β and A are unknown, we can get the estimates of them from the marginal distribution of y, and derive the cut-off points by the parametric bootstrap method(See Rao(2005)). The prediction interval is given by:

$$PI_i^{Syn}(t) = [x_i^{'}\hat{\beta} - t_1\sqrt{\hat{A}}, x_i^{'}\hat{\beta} + t_2\sqrt{\hat{A}}], t \in R,$$

where $\hat{\beta} = (X^{'}X)^{-1}X^{'}Y$ and $\hat{A}$ is the REML estimator.

In the above, (t_1, t_2) are obtained using

$$P^*[\theta_i^* < x_i^{'}\hat{\beta}^* - t_1\sqrt{\hat{A}^*}] = \alpha/2$$

$$P^*[\theta_i^* > x_i^{'}\hat{\beta}^* + t_2\sqrt{\hat{A}^*}] = \alpha/2$$

where the probability P^* is with respect to the parametric bootstrap distribution and θ^*, $\hat{\mu}^*$ and $\hat{A}^*$ are like θ, $\hat{\mu}$ and $\hat{A}$, except that bootstrap samples are used in place

of original sample. The bootstrap samples $(y_i^*, \theta_i^*), i = 1, \cdots, N$ are generated according to $\theta_i^* \overset{iid}{\sim} N[x_i'\hat{\beta}, \hat{A}]$ and $y_i^*|\theta_i^* \overset{ind}{\sim} N[\theta_i^*, D_i]$.

Since this method only use the second level of the model, the prior variance of θ_i is usually larger than the posterior variance of $\theta_i|Y$. The average length of the synthetic parametric bootstrap is always larger than the one based on the conditional distribution $\theta_i|Y$, which is used in CLL parametric bootstrap prediction interval. In addition, we can show that this prediction interval attains a coverage probability $1-\alpha$ with margin of error $O(m^{-1/2})$.

Theorem 4.1 *For (t_1, t_2) obtained by the above parametric bootstrap scheme, the following holds:*

$$P\{\theta_i \in [x_i'\hat{\beta} - t_1\sqrt{\hat{A}}, x_i'\hat{\beta} + t_2\sqrt{\hat{A}}]\} = 1 - \alpha + O(m^{-1/2}).$$

The coverage probability can be improved in Hall and Maiti(2006).

Method 3. CLL Parametric bootstrap method with REML estimator

Chatterjee, Lahiri and Li(2006) proposed a new parametric bootstrap method. They constructed the prediction interval using the EB(Empirical Bayes) estimator and its naive variance estimator. Their prediction interval for the Fay-Herriot model is given as:

$$PI_i^{CLL} = [(1-\hat{B})y_i + \hat{B}x_i'\hat{\beta} - t_1\sqrt{D_i(1-\hat{B})}, \\ (1-\hat{B})y_i + \hat{B}x_i'\hat{\beta} + t_2\sqrt{D_i(1-\hat{B})}]$$

where $\hat{B} = \hat{A}/(\hat{A} + D_i)$, $\hat{\beta} = (X'X)^{-1}X'Y$ and $\hat{A}$ is the REML estimator. The cut-off points (t_1, t_2) are computed from the bootstrap samples using:

$$P^*[\theta_i^* < (1-\hat{B}^*)y_i^* + \hat{B}^*x_i'\hat{\beta}^* - t_1\sqrt{D_i(1-\hat{B}^*)}] = \alpha/2$$

$$P^*[\theta_i^* > (1-\hat{B}^*)y_i^* + \hat{B}^*x_i'\hat{\beta}^* + t_2\sqrt{D_i(1-\hat{B}^*)}] = \alpha/2$$

where the probability P^* is with respect to the parametric bootstrap distribution and θ^*, $\hat{\beta}^*$ and $\hat{B}^*$ are like θ, $\hat{\beta}$ and $\hat{B}$, except that bootstrap samples are used in place of original sample. The bootstrap samples $(y_i^*, \theta_i^*), i = 1, \cdots, N$ are generated according to $\theta_i^* \overset{iid}{\sim} N[x_i'\hat{\beta}, \hat{A}]$ and $y_i^*|\theta_i^* \overset{ind}{\sim} N[\theta_i^*, D_i]$.

They also showed that this prediction interval has a coverage probability $1-\alpha$ with marginal error of $O(m^{-3/2})$. Hence, the CLL parametric bootstrap method is more accurate than the synthetic parametric bootstrap method.

Method 4: Parametric bootstrap method with ADM estimator

In this method, all the schemes to construct the prediction interval are the same as we used in method 3, except that we use ADM estimator of A instead of the REML estimator. When we use the REML method, it is possible to get negative value for the estimation of A, in which case we let $\hat{A} = 0$. Those zero estimates will make the variance estimation of θ_i problematic, and also cause trouble in the computing procedure. We will discuss it further in the simulation section.

5 Simulation Study

In this section, we compare four prediction intervals mentioned in the previous section using the simplest Fay-Herriot model with $D_i = 1$ and no covariates:

Level 1: $y_i|\theta_i \overset{ind}{\sim} N[\theta_i, 1], i = 1, \cdots, m;$

Level 2: $\theta_i \overset{ind}{\sim} N[\mu, A], i = 1, \cdots, m;$

In the simulation, we took $m = 20$. For each of the different values of $A = 0.2, 0.5, 1, 1.5$, we simulated $n = 10,000$ independent data sets $\{(y_i, \theta_i), i = 1, \cdots, 20\}$ from the above model. In each iteration we do the following steps to construct 4 prediction intervals for $\theta_i, i = 1, \cdots, m$:

1. Compute $\hat{\mu}$ and $\hat{A}$:

$$\hat{\mu} = \bar{y};$$
$$\hat{A}^{REML} = max(0, s^2 - 1),$$
$$s^2 = \sum(y_i - \bar{y})^2/(m-1);$$
$$\hat{A}^{ADM} = (\sum(y_i - \bar{y})^2 - m + 4 + \\ \sqrt{(m - 4 - \sum(y_i - \bar{y})^2)^2 + 8(m-2)})/(2m-4).$$

2. Generate bootstrap samples: $\theta_i^* \overset{ind}{\sim} N[\hat{\mu}, \hat{A}]$, $y_i^*|\theta_i^* \overset{ind}{\sim} N[\theta_i^*, 1]$, $i = 1, \cdots, m$, then get $\hat{\mu}^*$ and $\hat{A}^*$ by replacing original data y_i's in the formulae of step 1 with bootstrap samples y_i^*'s. Repeat this step for $N = 100,000$ times.

3. For each bootstrap sample, we can compute the pivot values for synthetic method and CLL method separately:

$$p^{Syn} = (\theta_i^* - \hat{\mu}^*)/\sqrt{\hat{A}^*};$$
$$p^{CLL} = (\theta_i^* - (1-\hat{B}^*)y_i^* - \hat{B}^*\hat{\mu}^*)/\sqrt{(1-\hat{B}^*)},$$
$$where\ \hat{B}^* = \hat{A}^*/(1+\hat{A}^*).$$

   For each method, those 100,000 pivot values can be used to build the empirical distribution, from which we can locate the two equal-tail $\alpha/2$ cut-off points (t_1, t_2).

4. Finally, we get the following prediction intervals:

$$PI_i^{Naive} = [y_i - z_{\alpha/2}, y_i + z_{\alpha/2}];$$
$$PI_i^{Syn} = [\hat{\mu} - t_1\sqrt{\hat{A}}, \hat{\mu} + t_2\sqrt{\hat{A}}];$$
$$PI_i^{CLL} = [(1-\hat{B})y_i + \hat{B}\hat{\mu} - t_1\sqrt{(1-\hat{B})}, \\ (1-\hat{B})y_i + \hat{B}\hat{\mu} + t_2\sqrt{(1-\hat{B})}].$$

Table 1: Average Coverage Probability of Prediction Intervals.

	A=0.2	A=0.5	A=1.0	A=1.5
Naive Method	0.9490	0.9503	0.9501	0.9500
Synthetic PB	0.6284	0.7979	0.9202	0.9531
CLL PB	0.6305	0.7988	0.9178	0.9495
CLL PB(ADM)	0.9698	0.9521	0.9461	0.9475
Prasad-Rao	0.8374	0.8484	0.897	0.9214

Table 2: Average Length of Prediction Intervals.

	A=0.2	A=0.5	A=1.0	A=1.5
Naive Method	3.92	3.92	3.92	3.92
Synthetic PB	2.12	3.25	4.54	5.34
CLL PB	1.76	2.53	3.20	3.44
CLL PB(ADM)	2.35	2.64	2.98	3.19
Prasad-Rao	1.68	2.16	2.70	3.01
counts of $\hat{A}^{REML} = 0$	3245	1472	369	109

We report the coverage probability(CP) and average length(AL) for the four prediction intervals. They are defined as:

$$CP_i = \sum [\theta_i \in PI_i]/n;$$
$$AL_i = \sum (length\ of\ PI_i)/n, i = 1, \cdots, m.$$

To save the space, we omit the detailed number for each small area, and only report the average number of 20 small areas.

The results showed that although the coverage probability of naive prediction interval can always attain the nominal value 0.95, the length of it is too large to draw any reasonable conclusion.

Comparing synthetic method and CLL method, they both apply the parametric bootstrap strategy to find the cut-off points, but based on different level of data. The former use the prior model only to derive the point estimator and its variance estimator, although its unknown parameters are estimated from the marginal distribution of y. The latter use the conditional distribution of $\theta_i|y_i$ to construct prediction interval by combining the information from the data and prior model. From table 1, we can see the coverage probabilities of those two method are almost same for different true values of A. However, the synthetic prediction interval have much larger average length than the CLL method. When $A = 1\ and\ 1.5$ its lengthes are even larger than the naive method. That may due to the fact that when A is large, the prior model is nor so reliable, in which case the composite estimator shows its superiority.

Next we look at the CLL method with the ADM estimator of A. The prediction interval constructed using this method is much better than the other three methods. Not only because its coverage probabilities are very closed to the nominal value 0.95, but also because its average lengthes are very small. Look at the last row of table 2, when A is small, the REML estimator is prone to produce negative estimator, correspondingly the prediction interval estimate does not perform well neither. The ADM estimator has no such problem, and it always gives positive estimate. This also suggests that when we use the REML method, we can truncate it to a small value like does the Stein estimator.

Acknowledgments

The author would like to thank Professor P. Lahiri for his helpful advisory and suggestions.

References

Butar, F. B., and Lahiri, P. (2002) "On measures of uncertainity of empirical Bayes small area estimators," to appear in *J. Statist. Plann. Inf.*

Chatterjee, S., Lahiri, P. and Li, H. (2006), On small area prediction interval problems, submitted to the *Annals of Statistics.*

Datta, G.S., Ghosh, M., Smith, D., and Lahiri, P. (1999), "On an asymptotic theory of conditional and unconditional coverage probabilities of empirical Bayes confidence intervals, "*Scand. J. Statist.* 29, 139-152.

Efron, B. (1979), "Bootstrap methods: another look at the jackknife, "*Ann. Statist.*, **7**, 101-118.

Fay, R. E. and Herriot, R. A. (1979)," Estimates of income for small places: an application of James-Stein procedure to census data,"*J. Amer. Statist. Assoc.*, **74**, 269-277.

Ghosh, M., and Rao, J.N.K. (1994),"Small Area Estimation: An Appraisal," *Statist. Sci.*, **9**, No. 1, 55-93, (with discussions).

Hall, P. and Maiti, T. (2006). On parametric bootstrap methods for small-area prediction. *J. Roy. Statist. Soc.* Ser. B, 68, 221-238.

Jiang, J., and Lahiri, P. (2006), Mixed model prediction and small area estimation, Editor's invited discussion paper, *Test,* Vol. 15, 1, 1-96.

Morris, C.N. (1983b)," Parametric empirical Bayes confidence intervals, " in *Scientific Inference, Data Analysis and Robustness*, Academic Press, New York, 25-50.

Rao, J. N. K. (2003), *Small Area Estimation*, Wiley, New York.

Tang, R.(2002), Fitting and Evaluating Certain Two-Level Hierarchical Models , PH.D. Thesis, Department of Statistics, Harvard University.

Wu, C. F. J. (1986) "Jackknife, bootstrap and other resampling methods in regression analysis," *Ann. Statist.*, **14**, 1261-1295.