A highly accurate simple small area confidence interval method

Masayo Yoshimori[1,3], Partha Lahiri[2]
[1] Graduate school of Engineering Science, Osaka University, Toyonaka, Japan
[2]University of Maryland, College Park, USA
[3]Corresponding author: masayo@sigmath.es.osaka-u.ac.jp

Abstract

We introduce a new adjusted residual maximum likelihood method in the context of producing an empirical Bayes confidence interval for a normal mean, a problem of great interest in different small area applications. Like other rival empirical Bayes confidence intervals such as the well-known parametric bootstrap method, the proposed interval is second-order correct. The proposed interval is carefully constructed so that it always produces an interval shorter than the corresponding maximum likelihood based direct confidence interval, a property not analytically proved for other competing methods. Moreover, the proposed method is not simulation-based and requires only a fraction of computing time needed for the parametric bootstrap confidence interval. A Monte Carlo simulation study demonstrates the superiority of the proposed method over other competing methods.

Keywords:Adjusted maximum likelihood, Coverage error, Empirical Bayes, Linear mixed model.

1 Introduction

Fay and Herriot (1979) considered empirical Bayes estimation of small area means θ_i using the following two-level Bayesian model and demonstrated, using real life data, that they outperform both the maximum likelihood and synthetic (e.g., regression) estimators.

The Fay-Herriot Model: For $i = 1, \cdots, m$,

Level 1 (sampling distribution): $y_i|\theta_i \overset{\text{ind}}{\sim} N(\theta_i, D_i)$;

Level 2 (prior distribution): $\theta_i \overset{\text{ind}}{\sim} N(x_i'\beta, A)$.

In the above model, level 1 is used to account for the sampling distribution of the maximum likelihood (direct) estimates y_i, which are weighted averages of observations from small area i. Level 2 prior distribution links the true small area means θ_i to a vector of $p < m$ known auxiliary variables $x_i = (x_{i1}, \cdots, x_{ip})'$, often obtained from various administrative records. The parameters β and A of the linking model are generally unknown and are estimated from the available data. As in other papers on the Fay-Herriot model (e.g., Datta et.al, 2005), the sampling variances D_i are assumed to be known. Note that the empirical Bayes estimator of θ_i obtained by

Fay and Herriot (1979) can be motivated as an empirical best prediction (EBP) estimator [in this case same as the empirical best linear unbiased prediction (EBLUP) estimator] of the mixed effect $\theta_i = x_i'\beta + v_i$, under the following linear mixed model:

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

where the v_i's and e_i's are independent with $v_i \overset{\text{iid}}{\sim} N(0, A)$ and $e_i \overset{\text{ind}}{\sim} N(0, D_i)$.

In this paper, we consider interval estimation of small area means θ_i. An interval, denoted by I_i, is called a $100(1-\alpha)\%$ interval for θ_i if $P(\theta_i \in I_i|\beta, A) = 1-\alpha$, $\forall \beta \in R^p, A \in R^+$, where the probability P is with respect to the Fay-Herriot model and R^+ is the positive part of the real line. Throughout the paper, $P(\theta_i \in I_i|\beta, A)$ is refereed to as the coverage probability of the interval I_i; that is, coverage is defined as the joint distribution of y and θ with fixed hyperparameters β and A. Most intervals proposed in the literature can be written as: $\hat{\theta}_i \pm q_\alpha \hat{\tau}_i(\hat{\theta}_i)$, where $\hat{\theta}_i$ is an estimator of θ_i, $\hat{\tau}_i(\hat{\theta}_i)$ is an estimate of the measure of uncertainty of $\hat{\theta}_i$ and q_α is suitably chosen in an effort to attain coverage probability close to the nominal level $1-\alpha$. Researchers have considered different choices for $\hat{\theta}_i$. For example, the choice $\hat{\theta}_i = y_i$ leads to the direct maximum likelihood based confidence interval I_i^D, given by

$$I_i^D : y_i \pm z_{\alpha/2}\sqrt{D_i},$$

where $z_{\alpha/2}$ is the upper $100(1-\alpha/2)\%$ point of $N(0,1)$. Obviously, for this direct interval, the coverage probability is $1-\alpha$. However, when D_i is large as in the case of small area estimation, it's length is too large to make any reasonable conclusion. We call an interval empirical Bayes confidence interval if we choose an empirical Bayes estimator for $\hat{\theta}_i$. We introduce the Bayesian credible interval in the context of the Fay-Herriot model. When the hyperparameters β and A are known, the Bayesian credible interval of θ_i is obtained using the posterior distribution of θ_i: $\theta_i|y_i \sim N[\hat{\theta}_i^B, \sigma_i(A)]$, where $\hat{\theta}_i^{\text{B}} \equiv \hat{\theta}_i^B(A) = (1-B_i)y_i + B_i x_i'\beta$, $B_i \equiv B_i(A) = \frac{D_i}{D_i+A}$, $\sigma_i(A) = \sqrt{\frac{AD_i}{A+D_i}}$ $(i = 1, \cdots, m)$. Such a credible interval is given by $I_i^B(A) : \hat{\theta}_i^B(A) \pm z_{\alpha/2}\sigma_i(A)$. The Bayesian credible interval cuts down the length of the direct confidence interval by $100 \times (1 - \sqrt{1-B_i})\%$ while maintaining the exact coverage $1-\alpha$ with respect to the joint distribution of y_i and θ_i. The maximum benefit from the Bayesian methodology is achieved when B_i is large, that is, when the prior variance A is much smaller than the sampling variances D_i.

Cox (1975) initiated the idea of developing an one-sided empirical Bayes confidence interval for θ_i for a special case of the Fay-Herriot model with $x_i^T\beta = \mu$ and $D_i = D$ $(i = 1, \cdots, m)$. The two-sided version of his confidence interval is given by:

$$I_i^{Cox}(\hat{A}) : \hat{\theta}_i^{EB}(\hat{A}) \pm z_{\alpha/2}\sigma(\hat{A}),$$

where μ is estimated by the sample mean $\bar{y} = m^{-1}\sum_{i=1}^m y_i$ and A by the ANOVA estimator: $\hat{A}_{ANOVA} = \max\left\{(m-1)^{-1}\sum_{i=1}^m (y_i - \bar{y})^2 - D,\, 0\right\}$.

In this paper, our goal is to find an empirical Bayes confidence interval of θ_i that (i) matches the coverage error properties of the best known empirical Bayes method such as the one proposed by Chatterjee et al. (2008), (ii) has length smaller than that of the direct method, and (iii) does not rely on simulation-based heavy computation. In section 2, we propose such a method by replacing the ANOVA method in the

Cox interval $I_i^{Cox}(\hat{A}_{ANOVA})$ by a carefully devised adjusted maximum likelihood estimator of A. In section 3, we compare our method with the direct, Cox and the parametric bootstrap method of Chatterjee et al. (2008).

2 A New Second-Order Efficient Empirical Bayes Confidence Interval

We call an empirical Bayes interval of θ_i second-order efficient if the coverage error is of order $o(m^{-1})$ and length less than that of the maximum likelihood based confidence interval. The goal of this section is to produce such an interval, which requires a fraction of computer time required by the recently proposed parametric bootstrap empirical Bayes confidence interval. Our idea is simple and involves replacement of the ANOVA estimator of A in the empirical Bayes interval proposed by Cox (1975) by a carefully devised adjusted residual maximum likelihood estimator of A.

We call $\hat{A}_{h_i}$ an adjusted residual maximum likelihood estimator of A if it maximizes the following adjusted residual likelihood:

$$L_{i;ad}(A) \propto h_i(A) \times L_{RE}(A),$$

with respect to A over $(0, \infty)$, where $h_i(A)$ is a general area specific adjustment factor; $L_{RE}(A)$ is the standard residual likelihood function given by

$$L_{RE}(A) = |X'V^{-1}X|^{-\frac{1}{2}}|V|^{-\frac{1}{2}} \exp\left(-y'Py/2\right);$$

$y = (y_1, \cdots, y_m)'$; $X' = (x_1, \cdots, x_m)$, $V = \text{diag}(A + D_1, \cdots, A + D_m)$;
$P = V^{-1} - V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}$.

We propose the following empirical Bayes interval of θ_i:

$$I_i^{Cox}(\hat{A}_{h_i}) : \hat{\theta}_i^{EB}(\hat{A}_{h_i}) \pm z_{\alpha/2}\sigma_i(\hat{A}_{h_i}),$$

where $\hat{\theta}_i^{\text{EB}} \equiv \hat{\theta}_i^{\text{EB}}(\hat{A}_{h_i}) = (1 - \hat{B}_i)y_i + \hat{B}_i x_i'\hat{\beta}$, $\hat{\beta} \equiv \hat{\beta}(\hat{A}) = (X'\hat{V}^{-1}X)^{-1}X'\hat{V}^{-1}y$;
$\hat{V} = \text{diag}(\hat{A}_{h_i} + D_1, \cdots, \hat{A}_{h_i} + D_m)$; $\hat{B}_i = D_i/(\hat{A}_{h_i} + D_i)$.

We obtain the following result under the following regularity conditions:
The adjustment factor $h_i(A)$ is free of y and four times continuously differentiable with respect with A over $[0, \infty)$, $rank(X) = p$ is fixed, $\sup_{i\geq 1} h_{ii} = O(m^{-1})$, where $h_{ii} = x_i'(X'X)^{-1}x_i$, and $0 < \inf_{i\geq 1} D_i \leq \sup_{i\geq 1} D_i < \infty$.

$$P\left\{\theta_i \in I_i^{Cox}(\hat{A}_{h_i})\right\} = 1 - \alpha + z\phi(z)\frac{a_i + b_i[h_i(A)]}{m} + O(m^{-\frac{3}{2}}),$$

where $a_i = -\frac{m}{tr(V^{-2})}\left[\frac{4D_i}{A(A+D_i)^2} + \frac{(1+z^2)D_i^2}{2A^2(A+D_i)^2}\right] - \frac{mD_i}{A(A+D_i)}x_i'Var(\tilde{\beta})x_i$,

$$b_i \equiv b_i[h_i(A)] = \frac{2m}{tr(V^{-2})}\frac{D_i}{A(A+D_i)} \times \tilde{l}_{i;ad},\ \tilde{\beta} = \hat{\beta}(A) = (X'V^{-1}X)^{-1}X'V^{-1}y,$$

where $\tilde{l}_{i;ad} = \frac{\partial \log(h_i(A))}{\partial A}$ and $z = z_{\alpha/2}$. Equation (2) suggests an area specific adjustment factor $h_i(A)$ that corrects the undercoverage of the Cox interval when ANOVA estimator for A is used. More specifically, for small area i, we suggest an adjusted REML estimator of A, where the adjustment factor satisfies the following differential

equation:

$$a_i + b_i\left[h_i(A)\right] = 0. \tag{1}$$

Let $\hat{A}_i$ denote the solution to (1). Then our proposed empirical Bayes confidence interval for θ_i is given by $I_i^{YL}(\hat{A}_i) : \hat{\theta}_i^{EB}(\hat{A}_i) \pm z_{\alpha/2}\sigma_i(\hat{A}_i)$. Since $\sigma_i(\hat{A}_i) < \sqrt{D_i}$, the length of this interval, like the original Cox interval $I_i^{Cox}(\hat{A}_{ANOVA})$, is always less than that of the direct interval I_i^D.

We now obtain a solution to (1). In the most general situation when a general least square (GLS) estimator of β in empirical Bayes (EB) estimator of θ_i is used, we have the following solution to (1):

$$h_i(A) = CA^{\frac{1}{4}(1+z^2)}(A+D_i)^{\frac{1}{4}(7-z^2)}\exp[\int \frac{1}{2}tr(V^{-2})x_i'(X'V^{-1}X)^{-1}x_i dA] \tag{2}$$

where $R = diag(D_1, \cdots, D_m)$ and C is a generic constant free of A. We use this generic constant C throughout the paper, which may have different expressions in different places. Additionally, we cannnot obtain the explisit $h_i(A)$. However, the calculation is only requested the deviation of $h_i(A)$, not explicit $h_i(A)$.

Note that when the ordinary least squares (OLS) estimator of β is used in the EB of θ_i, we cannot obtain an expression for the adjustment term $h_i(A)$ as a particular case of (2). In this case, we obtain the solution as:

$$h_i(A) = CA^{\frac{1}{4}(1+z^2)}(A+D_i)^{\frac{1}{4}(7-z^2)}\left[\prod_{i=1}^{m}(A+D_i)\right]^{\frac{1}{2}x_i'(X'X)^{-1}x_i} \exp\left[-tr(V^{-1})x_i'(X'X)^{-1}X'VX(X'X)^{-1}x_i/2\right]. \tag{3}$$

When $D_i = D$ $(i = 1, \cdots, m)$, the GLS of β is identical to the OLS. In this case, our solution for $h_i(A)$ can be considerably simplified to:

$$h_i(A) = CA^{\frac{1}{4}(1+z^2)}(A+D)^{\frac{1}{4}(7-z^2)+\frac{1}{2}mx_i'(X'X)^{-1}x_i}.$$

In this balanced case, we show the uniqueness of the solution $\hat{A}_i$ if $m > \frac{4+p}{1-x_i'(X'X)^{-1}x_i}$.

3 A Monte Carlo Simulation Study

In this section, we design a Monte Carlo simulation study to compare small sample performances of different confidence intervals of the small area means under the Fay-Herriot model. In subsection 3.1, we consider a Fay-Herriot model with a common mean as in Datta et al. (2005) and Chatterjee et al. (2008). In subsection 3.2, we consider a general Fay-Herriot model, where we use the sampling variances D_i and the auxiliary variables x_i from a real life data.

3.1 The Fay-Herriot Model with a common mean

Throughout this subsection, we assume a common mean $x_i'\beta = 0$, which is estimated using data as in other papers on small area estimation. Specifically, we generate $R = 10^4$ independent replicates $\{y_i,\ v_i,\ i = 1, \ldots, m\}$ using the following Fay Her-

riot model: $y_i = v_i + e_i$, where v_i and e_i are mutually independent with $v_i \overset{iid}{\sim} N(0, A)$, $e_i \overset{ind}{\sim} N(0, D_i),\ i = 1, \cdots, m$. We set $A = 1$. For the parametric bootstrap method, we consider $B = 6000$ bootstrap samples. In the unbalanced case, for $m = 15$, we consider five groups, say $G \equiv (G_1, G_2, G_3, G_4, G_5)$, of small areas, each with three small areas, such that the sampling variances D_i are the same within a given area. We consider the following two patterns of the sampling variances: (a) $(0.7, 0.6, 0.5, 0.4, 0.3)$ and (b) $(4.0, 0.6, 0.5, 0.4, 0.1)$. Note that in pattern (a) all areas have sampling variances less than A. In contrast, in pattern (b), sampling variances of all but one area are less than A. The patterns (a) and (b) correspond to the sampling variance patterns (a) and (c) of Datta et al. (2005).

The simulation results are displayed in Table 1. First note that while the direct method attains the nominal coverage most of the time it has the highest length compared to the other methods considered. The Cox method (denoted by COX.RE) cuts down the length of the direct method considerably at the expense of undercoverage, which is more severe for pattern (b) than pattern (a). This could be due to the presence of three outlying areas (i.e. with respect to the sampling variances) in G_1, which increases simulated probability of REML estimate of A being zero from 1% in pattern (a) to 8% in pattern (b). The parametric bootstrap method of Chatterjee et al. (2008) as implemented by Li and Lahiri (2010), denoted by CLL.LL, and our new method, denoted by COX.YL, perform very well in term of coverage although CLL.LL is showing a slight undercoverage. The CLL.LL is slightly better than ours in terms of length although the coverage of our method always exceeds the nominal coverage.

3.2 The Fay-Herriot model with D_i and x_i from real life data

In this subsection, we compare different interval estimators of small area means using D_i and x_i (census residuals) from the Small Area Income and Poverty Estimates (SAIPE) program of the U.S. Census Bureau. To save computational time (primarily because of the CLL.LL method), we only consider the 18 smallest areas (i.e., 18 states with the largest D_i) for the year 1999. The sampling variances D_i were obtained from a sampling error model of Otto and Bell (1995) that involved fitting a generalized variance function (GVF) to five years of direct variance and covariance estimates for each state produced by Fay and Train (1995). For further details on SAIPE, the readers are referred to Bell (1999) and the website: http://www.census.gov/hhes/www/saipe.html. We generate simulated data $\{(y_i, \theta_i),\ i = 1, \cdots, 18\}$ using the general Fay-Herriot model with x_i and D_i taken from the above-mentioned SAIPE data and using small $A = D_{\min}/10 \approx 1.1$, where $D_{\min}$ denotes the minimum D_i among the 18 states considered.

Again, the direct method attains the nominal coverage and the Cox method cuts down the length at the expense of severe undercoverage. The extent of undercoverage by the Cox method could be very severe (e.g., as low as about 35% for the smallest state). Our method corrects this undercoverage (our coverage is always more than the nominal coverage) while considerably reducing the length of the direct method. The CLL.LL method generally has lower length than ours, but it can have undercoverage when the leverage is large for the first area.

4 Acknowledgment

The first author's research was supported by JSPS KAKENHI Grant Number 242742. She conducted this research while visiting the University of Maryland, College Park, USA as a research scholar under the supervison of the second author whose research was supported in part by the National Science Foundation SES-085100.

Reference

Bell, W. R. (1999). "Accounting for uncertainty about variances in small area estimation", *Bulletin of the International Statistical Institute, 52nd Session, Helsinki, available at www.census.gov/hhes/www/saipe under "Publications."*.

Otto, M. C. and Bell, W. R. (1995). "Sampling Error Modeling of Poverty and Income Statistics for States", In *Proceedings of the Section on Government Statistics, American Statistical Association, Alexandria, VA.* 160-165.

Chatterjee, A., Lahiri, P. and Li, H. (2008). "Parametric bootstrap approximation to the distribution of EBLUP and related prediction intervals in linear mixed models", *Ann. Statist.* **36** 1221-1245.

Cox, D. R. (1975). "Prediction intervals and empirical Bayes confidence intervals", in:J.Gani(Ed.), *Perspectives in Probability and Statistics, Papers in Honor of M.S. Bartlett, Academic Press.* 47-55.

Datta, G. S., Rao, J. N. K. and Smith, D. D. (2005). "On measuring the variability of small area estimators under a basic area level model", *Biometrika* **92** 183-196.

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

Fay, R. E. and Train, G. F. (1997). "Small Domain Methodology for Estimating Income and Poverty Characteristics for States in 1993", In *Proceedings of the Social Statistics Section. American Statistical Association.* 183-188.

Li, H and Lahiri, P. (2010). "An adjusted maximum likelihood method for solving small area estimation problems", *J. Multivariate Anal.* **101** 882-892.

Table 1: Average Coverage Probability and Length: The Fay-Herriot Model with Common Mean

Pattern	G	Cox.RE		CLL.LL		Cox.YL		Direct	
a	1	89.8	(2.4)	94.5	(2.7)	95.3	(2.8)	95.1	(3.3)
	2	90.3	(2.3)	94.5	(2.5)	95.3	(2.6)	94.9	(3.0)
	3	90.6	(2.1)	94.6	(2.4)	95.2	(2.4)	95.2	(2.8)
	4	91.2	(2.0)	94.9	(2.2)	95.2	(2.2)	95.1	(2.5)
	5	91.1	(1.8)	94.3	(1.9)	95.0	(2.0)	94.7	(2.1)
b	1	88.3	(3.3)	94.5	(4.0)	95.8	(4.3)	94.9	(7.8)
	2	90.0	(2.3)	94.5	(2.5)	95.1	(2.6)	95.0	(3.0)
	3	90.4	(2.1)	94.6	(2.4)	95.3	(2.5)	94.9	(2.8)
	4	91.0	(2.0)	94.7	(2.2)	95.3	(2.2)	95.1	(2.5)
	5	93.1	(1.1)	94.7	(1.2)	95.0	(1.2)	95.0	(1.2)

Table 2: Average Coverage Probability and Length: The General Fay-Herriot Model using Auxiliary Variable and Sampling Variances from SAIPE Data

State	Ds	leverage	Cox.RE		CLL.LL		Cox.YL		Direct	
DC	28.2	0.63	35.1	(4.2)	92.7	(14.5)	95.3	(20.3)	94.6	(20.8)
DE	18.9	0.07	55.5	(4.0)	98.6	(9.4)	98.4	(11.0)	95.1	(17.1)
MS	17.9	0.08	53.3	(4.0)	98.2	(9.4)	97.4	(11.1)	94.3	(16.6)
LA	17.3	0.09	53.3	(3.9)	97.8	(9.5)	97.5	(11.1)	95.8	(16.3)
ME	16.3	0.12	51.7	(3.9)	97.6	(9.5)	97.4	(11.1)	94.9	(15.8)
MT	15.7	0.08	54.7	(3.9)	97.5	(9.2)	97.4	(10.7)	94.2	(15.5)
NM	14.7	0.06	54.5	(3.9)	98.5	(9.0)	98.3	(10.4)	95.5	(15.0)
MO	14.4	0.07	54.4	(3.9)	98.2	(9.0)	99.0	(10.4)	96.3	(14.9)
WV	14.2	0.06	55.4	(3.8)	98.1	(8.9)	97.4	(10.4)	94.8	(14.8)
RI	14.1	0.06	53.6	(3.8)	98.3	(8.9)	97.7	(10.4)	94.3	(14.7)
OR	13.6	0.06	55.4	(3.8)	97.2	(8.9)	97.3	(10.3)	93.8	(14.5)
ND	13.0	0.14	50.9	(3.8)	96.6	(9.2)	96.4	(10.7)	95.7	(14.1)
VT	12.9	0.14	50.6	(3.8)	96.4	(9.2)	96.1	(10.7)	94.6	(14.1)
SC	12.9	0.06	54.6	(3.8)	98.0	(8.7)	98.0	(10.1)	95.2	(14.1)
ID	12.7	0.08	54.0	(3.8)	97.7	(8.8)	97.2	(10.2)	94.4	(14.0)
AL	12.3	0.06	54.8	(3.8)	98.3	(8.6)	98.1	(10.0)	95.6	(13.7)
KS	12.0	0.08	53.6	(3.8)	97.6	(8.7)	97.3	(10.1)	94.6	(13.6)
GA	11.7	0.07	54.5	(3.7)	97.3	(8.7)	97.1	(10.0)	95.1	(13.4)