Small Area Interval Estimation

Partha Lahiri

Joint Program in Survey Methodology
University of Maryland, College Park

(Based on joint work with Masayo Yoshimori, Former JPSM Visiting PhD Student and Research Fellow of JSPS, Graduate School of Engineering Science, Osaka University)

June 12, 2013

The Fay Herriot Bayesian Model

Ref: Fay and Herriot (1979)

For $i = 1, \cdots, m$,

$$\textit{Level 1: (Sampling Distribution)}: \; y_i|\theta_i \sim N(\theta_i, D_i);$$
$$\textit{Level 2: (Prior Distribution)}: \; \theta_i \sim N(\mathbf{x}_i'\boldsymbol{\beta}, A)$$

where

- m : number of small area;
- y_i : direct survey estimate of θ_i;
- θ_i : true mean for area i;
- $\mathbf{x}_i$: $p \times 1$ vector of known auxiliary variables;
- D_i: known sampling variance of the direct estimate;
- The $p \times 1$ vector of regression coefficients $\boldsymbol{\beta}$ and model variance A are unknown.

Empirical Bayes (EB) Estimator of θ_i

Let $\hat{A}$ be a consistent estimator of A, for large m.

An EB of θ_i is given by

$$\hat{\theta}_i^{EB} = (1 - \hat{B}_i)y_i + \hat{B}_i \mathbf{x}_i' \hat{\boldsymbol{\beta}}.$$

where

- $\hat{B}_i = \frac{D_i}{\hat{A}+D_i}$
- $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\hat{A})$

Confidence Interval for 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 joint distribution of $\{(y_i, \theta_i), i = 1, \cdots, m\}$ under the Fay-Herriot model;
- R^+ is the positive part of the real line.

A General Form of Confidence Interval for θ_i

Most of the intervals proposed in the literature can be written as:

$$(\hat{\theta}_i + q_1(\alpha)\hat{\tau}_i(\hat{\theta}_i), \hat{\theta}_i + q_2(\alpha)\hat{\tau}_i(\hat{\theta}_i))$$

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 $\hat{\theta}_i$;
- $q_1(\alpha)$ and $q_2(\alpha)$ are chosen suitably in an effort to attain coverage probability close to the nominal level $1 - \alpha$.

Direct Confidence Interval

The choice $\hat{\theta}_i = y_i$ leads to the direct 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 $N(0,1)$.

Remarks:

- The coverage probability is $1-\alpha$;
- When D_i is large, the length is too large to make any reasonable conclusion.

Synthetic Confidence Interval

Ref: Hall and Maiti (JRSS, 2006)

$$(x_i^T \hat{\beta}_{(-i)} + q_1(\alpha)\sqrt{\hat{A}_{(-i)}},\ x_i^T \hat{\beta}_{(-i)} + q_2(\alpha)\sqrt{\hat{A}_{(-i)}})$$

where

- $\hat{\beta}_{(-i)}$ and $\hat{A}_{(-i)}$ are consistent estimators of β and A, respectively, based on all but the ith area data.
- $L_i^*[q_2(\alpha)] - L_i^*[q_1(\alpha)] = 1 - \alpha$ where L_i^* is a parametric bootstrap approximation of the distribution L_i of $\frac{\theta_i - x_i'\hat{\beta}_{(-i)}}{\hat{A}_{(-i)}}$.

Remarks:

- The coverage is $1 - \alpha + O(m^{-1.5})$.
- The method is synthetic (Rao 2005).
- This approach could be useful in situations especially when y_i is missing for the ith area.

Bayesian Credible Interval

Assume β and A are known.

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

where

- $\hat{\theta}_i^{\mathrm{B}} \equiv \hat{\theta}_i^B(A) = (1 - B_i)y_i + B_i x_i'\beta,$
- $B_i \equiv B_i(A) = \frac{D_i}{D_i + A},$
- $\sigma_i(A) = \sqrt{\frac{AD_i}{A+D_i}}$

Remarks:

- $\theta_i | y_i; \beta, A \sim N[\hat{\theta}_i^B(A), g_{1i} = \sigma_i^2(A)].$
- The Bayesian credible interval cuts down the length of the direct confidence interval by $100 \times (1 - \sqrt{1 - B_i})\%$
- The maximum benefit from the Bayesian methodology is achieved when B_i is large,

Hierarchical Bayesian Credible Interval

A review paper: Morris and Tang (2012)

- Assume prior on β and A, e.g.

$$\pi(\beta, A) \propto 1,\ \beta \in R^p,\ A \in R^+.$$

- Obtain the posterior distribution of $\theta_i | y = (y_1, \cdots, y_m)$ and use this posterior to construct credible interval for θ_i.
- Computation:
  - MCMC
  - Numerical Integration;
  - Laplace approximation.
  - ADM approximation

Empirical Bayes Confidence Interval

Ref: Cox (1975)

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

where

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

Remarks:

- Like the Bayesian credible interval, the length of the Cox interval is smaller than that of the direct interval.
- The distribution of $\frac{\theta_i - \hat{\theta}_i^{\mathrm{EB}}}{\sqrt{g_{1i}(\hat{A})}}$ is not a standard Normal. Thus, it is not appropriate to use the Normal quantile $z_{\alpha/2}$ as the cut-off points.
- The Cox empirical Bayes confidence interval introduces a coverage error of the order $O(m^{-1})$, not accurate enough in most small area applications.
- length of the interval is zero when $\hat{A}_{ANOVA} = 0$

Other EB Confidence Intervals

- Replace $z_{\alpha/2}$ by $z_{\alpha'/2}$ to reduce coverage error (Cox 1975).
- Replace $\sigma(\hat{A})$ by a measure of uncertainty that captures uncertainty due to estimation of the hyperparameters β and A (e.g., $\sqrt{g_{1i} + g_{2i} + 2g_{3i}}$) (Ref: Morris (1983) Prasad and Rao (1990))
- Replace $z_{\alpha/2}$ by $z_{\alpha/2} c_i(\hat{A})$ to reduce the coverage error to $O(m^{-1.5})$ (Datta et al. 2002; Basu et al. 2003; Yoshimori 2013)
- Parametric bootstrap (Laird and Louis 1987; Carlin and Louis 1996; Chatterjee et al. 2008)

Parametric Bootstrap Confidence Interval

Ref: Chatterjee, Lahiri and Li (AS, 2008)

- Use the distribution of $\frac{\theta_i^* - \hat{\theta}_i^{\mathrm{EB}*}}{\sigma_i(\hat{A}^*)}$ to approximate the distribution of $\frac{\theta_i - \hat{\theta}_i^{\mathrm{EB}}}{\sigma_i(\hat{A})}$.
- Compute $\hat{\beta}$ and $\hat{A}$;
- Draw bootstrap sample from the following bootstrap model:
  (i) $y_i^* | \theta_i^* \overset{ind}{\sim} N(\theta_i^*, D_i)$
  (ii)$\theta_i^* \overset{ind}{\sim} N(x_i'\hat{\beta}, \hat{A})$
- Compute $\hat{\beta}^*$ and $\hat{A}^*$ from y^*. Then we have $\hat{\theta}_i^{\mathrm{EB}*} = (1 - \hat{B}^*)y_i^* + \hat{B}^* x_i' \hat{\beta}^*$, and $\sigma_i^2(\hat{A}^*) = \frac{A^* D_i}{A^* + D_i}$;
- Compute $(\theta_i^* - \hat{\theta}_i^{\mathrm{EB}*})/\sigma_i(\hat{A}^*)$.

Remarks:

- Need strictly positive estimate of A (Li and Lahiri, JMVA 2010)
- In simulations, the parametric bootstrap based on EB performs better than the parametric bootstrap Hall-Maiti method based on synthetic estimator in terms of length (Yoshimori, 2013).

Parametric Bootstrap Confidence Interval

$$\mathrm{CI}_i^{\mathrm{PB}} = [\hat{\theta}_i^{\mathrm{EB}} + q_1\sigma_i(\hat{A}),\ \hat{\theta}_i^{\mathrm{EB}} + q_2\sigma_i(\hat{A})].$$

Theorem

Under reg. cond. $\Pr(\theta_i \in \mathrm{CI}_i^{\mathrm{PB}}) = 1 - \alpha + O(m^{-3/2})$,

A Research Question

Can we find an empirical Bayes confidence interval of θ_i that has the following properties?

Desired properties:

- coverage error of order $O(m^{-1.5})$;
- length smaller than that of the direct method;
- does not rely on simulation-based heavy computations.

A Generalization of Cox EB Confidence Interval

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

where

- $\hat{A}_{h_i}$ is obtained by maximizing the following adjusted residual likelihood:

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

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

A Higher-Order Expansion of Coverage

We obtain the following expansion under certain regularity conditions:

$$P(\theta_i \in I_i^{Cox}(\hat{A}_{h_i})) = 1 - \alpha + z\phi(z)\frac{a_i + b_i(h_i(A))}{m} + O(m^{-1.5}),$$

where

$$\begin{aligned} a_i &= -\frac{m}{tr(V^{-2})}\left[\frac{4D_i}{A(A+D_i)^2} + \frac{(1+z^2)D_i^2}{2A^2(A+D_i)^2}\right] - \frac{mD_i}{A(A+D_i)}x_i' Var(\tilde{\beta})x_i \\ b_i &= \frac{2m}{tr(V^{-2})}\frac{D_i}{A(A+D_i)} \times \frac{\partial \log(h(A))}{\partial A} \\ \tilde{\beta} &= \hat{\beta}(A) = (X'V^{-1}X)^{-1}X'V^{-1}y \end{aligned}$$

A Second-order Efficient Empirical Bayes Confidence Interval: Choice of $h_i(A)$

For small area i, we suggest an adjusted REML estimator of A where the adjustment factor satisfies the following differential equation:

$$a_i + b_i(h_i(A)) = 0.$$

Let $\hat{A}_i$ denote the solution to the above. 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 $\sigma_i(\hat{A}_i) < D_i$, 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.

Choice of $h_i(A)$ when OLS of β is used

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

where C is a generic constant free of A.
For the balanced case $D_i = D$ $(i = 1, \cdots, m)$

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

where C is a generic constant and free from A. In this balanced case, we show the uniqueness of the solution $\hat{A}_i$ if $m > \frac{4+p}{1-x_i'(X'X)^{-1}x_i}$.

Simulation Results: The Fay-Herriot Model with $x_i^T \beta = 0$ and Unequal Sampling Variances

Table : Coverage Probability and Average Length

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)

Simulation Results: Sampling Variances and Covariate from the 1999 SAIPE data

Table : Coverage Probability and Average length

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)