Jour. Ind. Soc. Ag. Statistics
49 (Golden Jub. No.), 1996-97 : 131–138

Estimating Small Domain Means on Borrowing Strength Across Time and Similar Domains —A Case Study

Arijit Chaudhuri and Tapabrata Maiti
Indian Statistical Institute, Calcutta, India

SUMMARY

A large survey population is supposed to be divisible into a number of non-overlapping parts called domains which vary in sizes. For several consecutive months it is assumed to change only a little and it is of interest to take simple random samples each month to estimate monthly domainwise means of a quantitative variable. As improvements on the respective sample means, available procedures that may exploit similarities of the domain-specific features and utilize past survey results, involve application of empirical Bayesian and Kalman filtering techniques. To facilitate their easy application drastic simplifying 'model postulation' is a common practice without applying any diagnostic check for its plausibility. Utilizing official records available in Indian Statistical Institute, Calcutta, it is demonstrated through a numerical exercise by simulation that the procedures seem to work well.

Key words : Empirical Bayes, Kalman filter, Models, Simple random sampling, Simulation, Small domains, Survey population.

1. Introduction

Ghosh and Meeden [1] have given the Empirical Bayes (EB) estimation procedure for estimating the means of disjoint domains, into which a finite population is divisible, on drawing a simple random sample without replacement (SRSWOR). Prasad and Rao [3] have given a procedure for estimating the mean square errors (MSE) of empirical best linear unbiased predictors (EBLUP) which are equivalent to EB estimators for small domain means. Meinhold and Singpurwalla [2] have given the Kalman filter (KF) procedure for estimation that gainfully employs past sample observations in a recursive way applying essentially a Bayesian approach. Applications of these techniques require postulating certain simplifying models. In large scale surveys it is not usually practicable to apply diagnostic checks for validity of such models. The purpose here is to show that adapting these procedures in suitably simple ways it is possible to derive appreciable improvements over simple domain-wise sample means as estimators of the respective domain means. The EB estimators and KF estimators along with their variance estimators are presented in Section 2. The numerical findings are presented in Section 3 to show how confidence intervals based on these procedures considerably outperform those based on

the sample means and the sample variances. The numerical data relate to 'dearness allowance' (DA) earned monthly by the workers of Indian Statistical Institute (ISI), Calcutta, who for administrative reasons are attached to several 'units', to be regarded as 'domains' in the present investigation. Data for 6 months from April through September in 1992 procured during another survey are used and cover only 15 'units' excluding the rest which are too small in size. Evaluation of the performances of the confidence intervals (CI) is according to several usual criteria described in Section 3.

2. *Empirical Bayes and Kalman Filter Estimators of Domain Means*

Let $U_t = (1_t, \ldots, i_t, \ldots, N_t)$ denote a survey population of size N_t at time t which takes the values $0, 1, \ldots, T (= 5)$ to respectively stand in the example above for the months of April, May, ..., September in 1992. Let U_{dt} denote the d th $(d = 1, \ldots, D)$ domain of U_t and U_{dt} be disjoint with $U_{d't}$ for $d \neq d'$ and for each t. Let s_t be an SRSWOR of size n_t from U_t and s_{dt} be the intersection of s_t with U_{dt}. Let $n_{dt} (\geq 2)$ be the cardinality of s_{dt}; $\bar{y}_{dt}$ the mean of a variable y of interest for the units of s_{dt} and v_{dt} the usual well-known variance estimator of $\bar{y}_{dt}$. The purpose is to estimate the 'domain means' $\bar{Y}_{dt}$. If a_{dt} is a point estimator of $\bar{Y}_{dt}$ admitting a positive-valued variance estimator b_{dt}, then it is usual to regard the pivot

$$c_{dt} = \frac{a_{dt} - \bar{Y}_{dt}}{\sqrt{b_{dt}}}$$

to be distributed as τ, the standard normal deviate having the distribution N(0, 1). Choosing α in (0, 1) and writing $\tau_{\alpha/2}$ as the $100\frac{\alpha}{2}\%$ point on the right tail area of the distribution of τ, the interval

$$a_{dt} \pm \tau_{\alpha/2} \sqrt{b_{dt}}$$

is supposed to provide a confidence interval (CI) with a nominal confidence coefficient $100 (1 - \alpha)$. One choice of (a_{dt}, b_{dt}) is obviously $(\bar{y}_{dt}, v_{dt})$. To find alternative but improved procedures, postulate first the following model under which one may write as follows, essentially in accordance with Ghosh and Meeden's [17] formulation:

(*i*) $\bar{y}_{dt} = \bar{Y}_{dt} + e_{dt}$;

(*ii*) e_{dt} for each fixed t is distributed as $N(0, V_{dt})$ independently across $d = 1, \ldots, D$; V_{dt} is assumed to be a known positive number;

(*iii*) $\overline{Y}_{dt} = \mu_t + \varepsilon_{dt}$ where μ_t is an unknown constant for each fixed t ;

(*iv*) ε_{dt} for each fixed t is distributed as $N(0, \sigma_t^2)$ independently across $d = 1, ..., D$ and σ_t (> 0) is an unknown constant ;

(*v*) for each d and each t, the random variable e_{dt} is distributed independently of ε_{dt}.

The following consequences, with obvious notations then follow :

(*a*) $\bar{y}_{dt} / \overline{Y}_{dt} \sim N(\overline{Y}_{dt}, V_{dt})$;

(*b*) $\overline{Y}_{dt} \sim N(\mu_t, \sigma_t^2)$;

(*c*) $\bar{y}_{dt} \sim N(\mu_t, \sigma_t^2 + V_{dt})$;

(*d*) The model-based covariance between $\bar{y}_{dt}$ and $\overline{Y}_{dt}$ is $C_m(\bar{y}_{dt}, \overline{Y}_{dt}) = \sigma_t^2$;

(*e*) $(\bar{y}_{dt}, \overline{Y}_{dt})' \sim N_2\left((\mu_t, \mu_t)', \begin{pmatrix} \sigma_t^2 + V_{dt} & \sigma_t^2 \\ \sigma_t^2 & \sigma_t^2 \end{pmatrix}\right)$ and

(*f*) $\overline{Y}_{dt}/\bar{y}_{dt} \sim N\left(\mu_t + \dfrac{\sigma_t^2}{\sigma_t^2 + V_{dt}} (\bar{y}_{dt} - \mu_t), \dfrac{\sigma_t^2 V_{dt}}{\sigma_t^2 + V_{dt}}\right)$

Let

$$\tilde{\mu}_t = \frac{\Sigma_d \dfrac{\bar{y}_{dt}}{\sigma_t^2 + V_{dt}}}{\Sigma_d \dfrac{1}{\sigma_t^2 + V_{dt}}}$$

denoting by Σ_d the sum over the D domains. Then its model-based variance is

$$V_m(\tilde{\mu}_t) = \Sigma_d \frac{1}{\sigma_t^2 + V_{dt}}$$

Taking V_{dt} as v_{dt} solving iteratively the equation

$$\Sigma_d \frac{(\bar{y}_{dt} - \tilde{\mu}_t)^2}{\sigma_t^2 + v_{dt}} = D - 1$$

for σ_t^2, we propose estimating σ_t^2 by $\hat{\sigma}_t^2$, applying the method of moments. Substituting this $\hat{\sigma}_t^2$ in $\tilde{\mu}_t$, denote the latter by $\hat{\mu}_t$ and repeating this in $\gamma_{dt} = \dfrac{\sigma_t^2}{\sigma_t^2 + v_{dt}}$ and in $V_m(\tilde{\mu}_t)$, denote them respectively by $\hat{\gamma}_{dt}$ and W_t.

Then, as is well-known,

$$m_{dt} = \hat{\gamma}_{dt}\,\bar{y}_{dt} + (1 - \hat{\gamma}_{dt})\,\hat{\mu}_t$$

is treated as the EBE and equivalently the EBLUP for $\bar{Y}_{dt}$. Adopting Prasad and Rao's [3] method a variance estimator for m_{dt} is

$$\hat{M}_{dt} = g_{1dt}(\hat{\sigma}_t^2) + g_{2dt}(\hat{\sigma}_t^2) + 2g_{3dt}(\hat{\sigma}_t^2)$$

where

$$g_{1dt}(\hat{\sigma}_t^2) = \hat{\gamma}_{dt} v_{dt},\ g_{2dt}(\hat{\sigma}_t^2) = \frac{(1-\hat{\gamma}_{dt})^2}{\sum_d \dfrac{1}{\hat{\sigma}_t^2 + v_{dt}}}$$

$$g_{3dt}(\hat{\sigma}_t^2) = \frac{v_{dt}}{(\hat{\sigma}_t^2 + v_{dt})^3}\,\bar{V}(\hat{\sigma}_t^2),\ \bar{V}(\hat{\sigma}_t^2) = \frac{2}{D^2}\sum_d (\hat{\sigma}_t^2 + v_{dt})$$

So far we have sought to borrow strength only across domains utilizing data relevant to a given time point. Next to proceed to utilize past data we need to postulate time series models in the following simple way taking the clue from Meinhold and Singpurwalla [2] so as to be able to write as follows, taking $\bar{M}_{dt}$, $\bar{W}_t$ as known positive numbers :

(1) $m_{dt} = \mu_t + \eta_{dt}$, $\eta_{dt} \sim N(0, \bar{M}_{dt})$, 'independently over t and d'; $\mu_t = \mu_{t-1} + w_t$, $w_t \sim N(0, \bar{W}_t)$, 'independently over t' and also independently of η_{dt} for each t and d; $t = 1, \ldots, T$.

(2) $\mu_0 \sim N(m_{d0}, \bar{M}_{d0})$

(3) η_{dt} and w_t are both distributed 'independently' or μ_{t-1} for $t = 1, \ldots, T$.

Writing $\Delta_{d1} = m_{d1} - m_{d0}$ and $\bar{R}_{d1} = \bar{W}_1 + \bar{M}_{d0}$; note the following consequences :

(*A*) $\Delta_{d1} / m_{d0} \sim N(0, \bar{R}_{d1} + \bar{M}_{d1})$

(*B*) $\mu_1 / m_{d0} \sim N(m_{d0}, \bar{R}_{d1})$

(*C*) The model-based covariance between Δ_{d1} and μ_1 conditional on a given m_{d0} is $C_m(\Delta_{d1}, \mu_1)/m_{d0} = \overline{R}_{d1}$,

(D) $(\Delta_{d1}, \mu_1)' \sim N_2\left((0, m_{d0})', \begin{pmatrix} \overline{R}_{d1} + \overline{M}_{d1} & \overline{R}_{d1} \\ \overline{R}_{d1} & \overline{R}_{d1} \end{pmatrix}\right)$ and

(*E*) μ_1 / Δ_{d1}, i.e.

$$\mu_1 / (m_{d0}, m_{d1})' \sim N\left(m_{d0} + \frac{\overline{R}_{d1}}{\overline{R}_{d1} + \overline{M}_{d1}} \Delta_{d1},\ \overline{R}_{d1} - \frac{\overline{R}_{d1}^2}{\overline{R}_{d1} + \overline{M}_{d1}}\right).$$

So, recalling the relation (iii) above in our postulated model it is reasonable to estimate $\overline{Y}_{d1}$ by

$$\mu_{d1}^* = \frac{R_{d1}}{R_{d1} + \hat{M}_{d1}} m_{d1} + \frac{\hat{M}_{d1}}{R_{d1} + \hat{M}_{d1}} m_{d0}$$

Here we have replaced $\overline{M}_{d1}$ by $\hat{M}_{d1}$, $\overline{W}_1$ by W_1 and written $R_{d1} = W_1 + \hat{M}_{d0}$ in the expression for the posterior expectation of μ_1 given (m_{d0}, m_{d1}), to derive μ_{d1}^*. As a variance estimator for it, take the naive formula

$$\Sigma_{d1}^* = \frac{R_{d1} \hat{M}_{d1}}{R_{d1} + \hat{M}_{d1}}$$

which though an under estimator is simple and works well with the example.

Proceeding recursively, for $t = 2, ..., T$ we similarly get Kalman filter estimators μ_{dt}^* for $\overline{Y}_{dt}$ along with variance estimators E_{dt}^* with obvious notations. To save space, suppress the formulae easy to derive extending the results (A) – (E) to cover $t = 2, ..., T$; one may consult Meinhold and Singpurwalla [2] for details.

Our main purpose is to examine the relative efficacies of $100(1 - \alpha)$ % CI given by

$$\text{(I) } \bar{y}_{dt} \pm \tau_{\alpha/2} \sqrt{v_{dt}},\ \text{(II) } m_{dt} \pm \tau_{\alpha/2} \sqrt{\hat{M}_{dt}},\ \text{(III) } \mu_{dt}^* \pm \tau_{\alpha/2} \sqrt{\Sigma_{dt}^*}$$

utilizing the numerical data available from official records in Indian Statistical Institute (ISI), Calcutta.

3. *Evaluation of Confidence Intervals by a Simulation Study*

In April 1992 from the pay roll of ISI, Calcutta we got hold of 876 workers who belonged to 15 different mutually exclusive administrative 'units' and continued to remain within those units upto September 1992. Each month we

independently took a sample of 200 of these workers using SRSWOR scheme. The variable of interest y in this study is taken as the worker's dearness allowance (DA) earned in the month. We intend to estimate each month the average DA earned by the workers in the respective 'units' which are treated as the 'domains' which vary in size from 29 to 125. Then, taking $\alpha = 0.05$ we construct the CI's by formulae (I)-(III). To examine their performances we replicate the samples R = 1000 times. For comparison we apply the usual criteria $C_1 - C_6$ described below.

By Σ_r we denote sum over the replicates and consider the CI's calculated using pairs (a_{dt}, b_{dt}), (a'_{dt}, b'_{dt}) where a_{dt}, a'_{dt} are estimators/predictors for $\bar{Y}_{dt}$ and b_{dt}, b'_{dt} their positive-valued variance estimators. The criteria are :

C_1 : ACP (Actual coverage percentage) $\equiv$ the percent of replicates for which the CI covers $\bar{Y}_{dt}$- the closer it is to 95 the better, everything else remaining in tact.

C_2 : ACV (Average coefficient of variation) $\equiv \frac{1}{R} \Sigma_r \frac{\sqrt{b_{dt}}}{a_{dt}}$ – this reflects the length of the CI.

C_3 : ARE (Average relative error) $\equiv \frac{1}{R} \Sigma_r \left| \frac{a_{dt} - \bar{Y}_{dt}}{\bar{Y}_{dt}} \right|$.

C_4 : PMSE (Pseudo mean square error) $\equiv \frac{1}{R} \Sigma_r (a_{dt} - \bar{Y}_{dt})^2$

$= \text{PMSE}(a_{dt})$.

C_5 : PCV (Pseudo coefficient of variation) $\equiv \frac{1}{\bar{b}_{dt}} \sqrt{\frac{1}{R} \Sigma_r (b_{dt} - \bar{b}_{dt})^2}$

where $\bar{b}_{dt} = \frac{1}{R} \Sigma_r b_{dt}$.

C_6 : RE (Relative efficiency) of a_{dt} versus $a'_{dt} \equiv \frac{\text{PMSE}(a'_{dt})}{\text{PMSE}(a_{dt})} 100$.

The smaller the magnitudes of $C_2 - C_5$ the better the choice (a_{dt}, b_{dt}). The larger the value of C_6 the better a_{dt} relative to a'_{dt}.

The table below presents for illustration, numerical values of our findings relating to ISI, Calcutta about the criteria for procedures (I)-(III) with $\alpha = 0.05$ for selected domains for 2 months.

Table. Findings on procedures (I)-(III) and relative efficiencies of m_{dt}, μ^*_{dt} are given versus $\bar{y}_{dt}$, downwards in succession. Values for September are given following slashes after those for June.

Domain size	ACP	10^5 ACV	10^5ARE	10^3PCV	RE
29	85.5/86.4	12968/12659	13/5	1071/1028	
	86.7/87.2	12039/11838	12/2	809/766	113.24/118.64
	86.5/87.5	8868/8713	12/5	436/410	217.10/254.25
52	87.2/89.0	10435/10122	9/10	840/654	
	87.8/89.5	9995/9738	9/10	695/602	119.09/106.76
	87.7/90.1	7733/7633	3/9	422/400	193.09/197.88
69	89.7/92.3	12032/8711	3/21	565/537	
	91.0/92.7	11464/8479	1/21	507/508	120.17/103.81
	91.8/94.0	8752/7032	0/21	322/346	257.39/141.28
71	89.8/90.0	8095/8968	5/20	666/644	
	90.3/90.6	7896/8709	6/19	628/601	104.52/107.55
	91.9/91.2	6598/7010	5/9	408/371	188.51/207.17
95	92.6/90.0	8086/7923	9/2	492/479	
	93.1/90.5	7937/7782	9/2	460/449	116.50/114.33
	94.1/92.9	6372/660	8/4	274/287	222.09/218.32
125	93.6/91.2	3175/3327	1/0	417/489	
	93.8/91.9	3163/3314	1/0	416/487	98.51/97.66
	94.2/93.0	3078/3213	1/0	392/449	108.86/111.15

Concluding Remarks : Though the postulated models seem too simple and not quite realistic because correlations among domain-specific values and possible serial correlations in the error terms in the time series model are ignored and no tests for validity are applied to check the models, our main purpose is well served. We intended to achieve improvements upon the basic confidence interval

$$\bar{y}_{dt} \pm \tau_{\alpha/2} \sqrt{v_{dt}}$$

through borrowing strength across domains and time by modelling. Undoubtedly improvements are quite palpable from the simulated study presented in the table. Further possibilities for improvements by dint of more sophisticated procedures

like hierarchical Bayesian methods employing Markov Chain Monto Carlo techniques and Gibbs sampling methods are not tried here to present simplicity which is a necessity often in practice in large-scale surveys.

REFERENCES

[1] Ghosh, M. and Meeden, G., 1986. Empirical Bayes estimation in finite population sampling. *J. Amer. Statist. Assoc.*, **81**, 1058-1062.

[2] Meinhold, R.J. and Singpurwalla, N. D., 1983. Understanding the Kalman filter. *Amer. Stat.*, **37**, 2, 123-127.

[3] Prasad, N.G.N. and Rao, J.N.K., 1990. The estimation of mean squared errors of small area estimators. *J. Amer. Statist. Assoc.*, **85**, 163-171.