SMALL DOMAIN ESTIMATION BY EMPIRICAL BAYES AND KALMAN FILTERING PROCEDURES - A CASE STUDY

ARIJIT CHAUDHURI, ARUN KUMAR ADHIKARY AND
ARUP KUMAR SEAL
INDIAN STATISTICAL INSTITUTE, 203,B.T.ROAD,
CALCUTTA-700035, INDIA.

Abstract

An application of empirical Bayes and Kalman filtering techniques is reported, using live data from Indian Statistical Institute (ISI), Calcutta , to illustrate how initial small domain estimators may be vastly improved upon. A stratified two stage sampling procedure is adopted, allowing selection of first stage units with unequal probabilities but of second stage units with equal probabilities. Standard design-based estimators for domain totals are initialized based on domain specific survey data alone. Strength is then borrowed across domains and from past surveys. The resulting gains in efficacy are numerically demonstrated, through replicated sampling from official records.

1. Introduction

We consider sampling from a survey population to estimate totals of a variable of interest for several non-overlapping domains of different sizes. For improved small domain estimation, borrowing strength 'across similar

domains' and 'from past sample observations' respectively by 'empirical Bayes' and 'Kalman filtering' techniques is a usual pratice. As model postulation is involved in applying these corrective techniques it is of interest to examine how the methods apply in practice. We bear in mind that sophisticated diagnostic tests are rarely applied in large-scale surveys. So, we undertake a case study presented below in brief. Owing to paucity of resources no survey could be undertaken but only easily accessible official records from our institute are utilized. Consequently the variables used for estimation and sample selection are not very realistic. Yet the efficacy of the two techniques noted above is well-illustrated with our limited empirical exercise. We initiate with estimators of domain totals that use only domain-specific sample observations for the respective time points. Then we check that empirical Bayes estimators that borrow strength across other domains fare better. Finally we find that Kalman filter estimators that moreover use past data are even better. The models we postulate to apply these two techniques may appear too simplistic and not quite realistic. But the reason why we persist with them is that even with their limitations they yield fruitful results; with further refinements in the models possibly even better results may be forthcoming. But for simplicity and to derive quick results we do not explore more sophisticated models and techniques. In our view these techniques should be employed in practice in similar situations with prospects for good results.

2. Formulation of the problem and the method of solution

About 1200 workers, of ISI, Calcutta for adminstrative reasons are attached to several different 'units'. We consider an 'artificial' problem of classifying them into several 'domains' in order of 'monthly take-home pay' of a worker and estimating the respective 'domain totals' of dearness allowances (DA) "earned by the workers in the aggregate" of respective domains on taking a sample of 200 workers from the institute as a whole. For sampling, the ISI 'units' are stratified as follows. Our available monthly data are for April through September, 1992. Every 'unit' with 50 or more workers in 'April 1992' is supposed to constitute a separate stratum and there are 9 such strata; those with number of workers between 25 and 49 form the 10th stratum which has 10 "units" and the remaining 'units' totalling 20 together give the 11th stratum. Formation of strata is thus with reference to April 1992 figures alone and is persisted with for the subsequent months. From the first 9 strata a sample of 120 workers is chosen with a propor-

tional allocation of sample-size. The workers for these respective strata are chosen employing the Rao, Hartley and Cochran (RHC,1962) scheme, narrated in appendix below, with basic pay of worker as the 'size-measure' for sample selection. From the 10th stratum of 10 'units' a sample of 5 'units' is chosen by RHC scheme and from the 11th stratum of 20 'units' a sample of 10 'units' is chosen again by RHC scheme taking 'number of workers in the unit' as the size-measure for sample selection. From each of the first set of selected 5 'units' a simple random sample (SRS) of 12 workers is chosen without replacement (WOR) and from each of the second set of 10 'units' an SRSWOR of 2 workers is chosen. This gives a total sample of 200 workers. For each of the 6 months, April-September in 1992, the same sampling scheme is followed; of course the total size N of the population and strata sizes $N_h(h = 1, \cdots, H = 11)$ vary across the months. Domain sizes N_d vary appreciably from 2 to 156 and so of course the domain-wise sample-sizes n_d; $d = 1, \cdots, D$. The number of domains D turns out 25. Let y_i be the monthly DA of the ith worker, $i = 1, \cdots, N$ and $I_{di} = 1$ or 0 according as the ith worker belongs to the dth domain or else. We first derive an estimate $\hat{Y}_{dh}$ of the 'stratum-total' Y_{dh} of the domain specific values of $y_{di} = y_i I_{di}$ for every stratum $h = 1, \cdots, H$ and every domain $d = 1, \cdots, D$. Then aggregating the estimates across the strata we derive the respective estimated domain totals, which we denote by t_d $(d = 1, \cdots, D = 25$). In the "Appendix" we indicate a formula for t_d and for a variance estimator v_d of t_d. Since $t_d = \sum_{h=1}^{H} \hat{Y}_{dh}$, a formula for v_d is obtained by summing the variance estimators of $\hat{Y}_{dh}$, over $h = 1, \cdots, H$; formulae for the latter are given in the Appendix.

To improve upon t_d, we apply as follows an empirical Bayes procedure. Let x denote the 'gross pay' of a worker and X_d its total for domain d. Following the standard literature, noting Prasad and Rao (1990), Ghosh and Rao (1994) among others, we postulate the following :

$t_d = Y_d + e_d$ such that $e_d \sim N(0, v_d)$ i.e. e_d is distributed 'independently' normally with mean 0 and variance v_d, for $d = 1, \cdots, D$; further, let

$Y_d = \beta X_d + \epsilon_d$ such that β is an unknown regression coefficient and $\epsilon_d \sim$ N(0,A), i.e. ϵ_d is distributed "independently of each other and of e_d for every d", with A as an unknown positive quantity. Here a common slope across the domains is assumed in oder to borrow strength from similar domains. Independence assumption is for simplicity of resulting procedures. It follows that (t_d , Y_d) has a bivariate normal distribution with mean vector $(\beta X_d , \beta X_d)$ and the dispersion matrix

$$\begin{pmatrix} A + v_d & A \\ A & A \end{pmatrix}.$$

So the "posterior" distribution of Y_d given t_d is normal with mean $t'_d = \frac{A}{A+v_d} t_d + \frac{v_d}{A+v_d} \beta X_d$, the Bayes estimator of Y_d and variance $\frac{Av_d}{A+v_d}$.

Let

$$\frac{\sum_{d=1}^{D} t_d X_d/(A + v_d)}{\sum_{d=1}^{D} X_d^2/(A + v_d)} = \tilde{\beta}.$$

Then $\sum_{d=1}^{D}(t_d - \tilde{\beta}X_d)^2/(A+v_d)$ follows Chi-square distribution with (D-1) degrees of freedom. Applying the method of moments, by iteration we solve for A, the equation

$$\sum_{d=1}^{D}(t_d - \tilde{\beta}X_d)^2/(A + v_d) = D - 1$$

to find an estimator $\hat{A}$ for A. Then follows the empirical Bayes estimator (EBE), writing $\hat{\beta}$ for $\tilde{\beta}$ with A replaced by $\hat{A}$ in $\tilde{\beta}$,

$$m_d = \frac{\hat{A}}{A+v_d} t_d + \frac{v_d}{A+v_d} \hat{\beta}X_d$$

for Y_d. Following Prasad and Rao (1990) we estimate its mean square error (MSE) by

$$\hat{M}_d = g_{1d}\left(\hat{A}\right) + g_{2d}\left(\hat{A}\right) + 2g_{3d}\left(\hat{A}\right)$$

writing $\gamma_d = \frac{A}{A+v_d}$, $g_{1d}(A) = \gamma_d v_d$, $g_{2d}(A) = (1 - \gamma_d)^2 X_d^2/\sum_{d=1}^{D} \frac{X_d^2}{A+v_d}$, $g_{3d}(A) = \frac{v_d^2}{(A+v_d)^3} \bar{V}(A)$, $\bar{V}(A) = \frac{2}{D^2}\sum_{d=1}^{D}(A + v_d)^2$. Also $g_{jd}(\hat{A})$ is the value of $g_{jd}(A)$ with A replaced by $\hat{A}$ for j=1,2,3. Further we note the model-based variance of $\tilde{\beta}$ namely $V_m(\tilde{\beta}) = \frac{1}{\sum_{d=1}^{D} X_d^2/(A+v_d)}$ and estimate it by $W = \frac{1}{\sum_{d=1}^{D} X_d^2/(\hat{A}+v_d)}$.

For a further improvement upon m_d using "past data", we apply as follows the Kalman filtering technique as given, among others, by Meinhold and Singpurwalla (1983) . By $t = 0, 1, \cdots, T$ with T as 5, let us denote the successive months April,May,...,September in 1992 and attach this time subscript t to previous symbols $m_d, \beta, X_d, \hat{M}_d, W$ etc with obvious implications. Let us introduce further modelling to write $(i) m_{dt} = \beta_t X_{dt} + \eta_{dt}$, $(ii) \eta_{dt} \sim N(0, \hat{M}_{dt})$, $(iii) \beta_t = \beta_{t-1} + \xi_t$ and postulate for simplicity that $(iv) \eta_{dt}$ is independent of ξ_t, $(v) \xi_t \sim N(0, W_t)$, where $W_t = \frac{1}{\sum_{d=1}^{D} X_{dt}^2/(\hat{A}_t+v_{dt})}$, $(vi) \beta_0 \sim N(\phi_{d0}, \Sigma_{d0})$, where $\phi_{d0} = \frac{m_{d0}}{X_{d0}}$, $\Sigma_{d0} = \frac{\hat{M}_{d0}}{X_{d0}^2}$ using (i) above. Then the recursive steps of Kalman filtering are as follows :

Conditional on m_{d0}, the distribution of β_1 is normal $N(\phi_{d0}$, $R_{d0} = \Sigma_{d0} + W_1)$; let $m^*_{d1} = X_{d1}\,\phi_{d0}$, $\Delta_{d1} = m_{d1} - m^*_{d1} = X_{d1}\,(\beta_1 - \phi_{d0}) + \eta_{d1}$; then the conditional distribution, given m_{d0} , of $(\Delta_{d1}, \beta_1)'$ is bivariate normal with mean vector $(0, \phi_{d0})'$ and dispersion matrix

$$\begin{pmatrix} X^2_{d1}\,R_{d1} + \hat{M}_{d1} & X_{d1}\,R_{d1} \\ X_{d1}\,R_{d1} & R_{d1} \end{pmatrix} .$$

From this follows that the conditional distribution of β_1/Δ_{d1} i.e. of β_1 given (m_{d0}, m_{d1}) , is normal with mean

$$\frac{X_{d1}R_{d1}}{X^2_{d1}R_{d1} + \hat{M}_{d1}}\, m_{d1} + \frac{\hat{M}_{d1}}{X^2_{d1}R_{d1} + \hat{M}_{d1}}\, \phi_{d0} = \beta^*_1 \text{ , say ,}$$

and variance $\frac{R_{d1}\hat{M}_{d1}}{R_{d1}X^2_{d1} + \hat{M}_{d1}} = V_1$, say. Then the Kalman filter (KF) estimator of Y_{d1} is taken as $K_{d1} = X_{d1}\,\beta^*_1$ and its measure of error as the estimated MSE namely $X^2_{d1}\,V_1 = M_d(1)$, say. This procedure is repeated to derive KF estimators K_{dt} for Y_{dt} ($t = 2, \cdots, T$) and their estimated MSE's $M_d(t)$ in an obvious way. One may consult Meinhlod and Singpurwalla (1983).

We expect m_{dt} to improve upon the initial t_{dt} and K_{dt} upon m_{dt}. To examine this theoritically is difficult. So, we examine the relative performances of 95% confidence intervals (CI) for Y_{dt} respectively given by $(i)t_{dt} \pm 1.96\,\sqrt{v_{dt}}$, $(ii)m_{dt} \pm 1.96\,\sqrt{\hat{M}_{dt}}$ and $(iii)K_{dt} \pm 1.96\,\sqrt{M_d(t)}$ based on the usual assumption of normality of the distribution of a pivot like

$$\frac{\hat{Y}_{dt} - Y_{dt}}{\sqrt{\hat{M}(\hat{Y}_{dt})}} ,$$

writing $\hat{Y}_{dt}$ for an estimator of Y_{dt} and $\hat{M}(.)$ for its MSE estimator. For this we report a numerical exercise, carrying out a simulation, with R = 10,000 replicates of samples drawn in a manner described already, calculating the above three CI's I - III.

The following criteria are considered for evaluation of relative performances of the CI's constructed as (I)-(III), writing $\sum_r$ as sum over the replicated samples.

(1.) *ACP* (Actual coverage percentage) $\equiv$ the percent of replicates for which a CI covers Y_{dt} - the closer it is to 95, other things remaining in tact, the better.

2. ACV (Average coefficient of variation) $\equiv \frac{1}{R} \sum_r \frac{\sqrt{\hat{M}(.)}}{\hat{Y}_{dt}}$; the smaller it is the better as it reflects the length of CI;

3. RE (Relative efficiency). Define

$$PMSE\left(\hat{Y}_{dt}\right) = \frac{1}{R} \sum_r \left(\hat{Y}_{dt} - Y_{dt}\right)^2$$

(a) RE of m_{dt} vs t_{dt} is $\frac{PMSE(t_{dt})}{PMSE(m_{dt})}$ and (b) similarly of K_{dt} vs t_{dt}, (c) K_{dt} vs m_{dt}; the larger the values of (a), the greater the advantage with EB method; likewise for (b) - (c).

We present the numerical findings in 2 tables in the next section, selectively for a few domains and for the 2 months of June and September, 1992.

3. Numerical Findings on Comparative Performances of Procedures

Values are given for t_{dt}, m_{dt} and K_{dt} successively, separted by commas.

Table 1
Relative performances for the month of June

Domain size	Domain total(Rs.)	ACP			$10^2 ACV$			RE		
		I	II	III	I	II	III	(a)	(b)	(c)
60	44213.13	88.5,	99.0,	98.4	29.5,	29.1,	19.1	6.7,	12.6,	1.9
117	83787.00	71.0,	99.2,	99.6	20.1,	17.8,	11.3	12.6,	40.1,	3.2
154	119838.70	76.4,	99.1,	99.6	16.5,	13.6,	9.7	11.3,	26.7,	2.4
143	125171.00	85.7,	99.4,	99.7	18.8,	13.8,	9.3	13.6,	40.8,	3.0
115	117177.00	91.6,	98.4,	99.0	21.4,	14.7,	10.1	7.6,	19.5,	2.6
103	119669.00	89.9,	97.3,	97.0	23.0,	15.0,	10.5	6.5,	11.0,	1.7
115	142675.11	90.3,	97.7,	97.9	20.7,	13.2,	9.6	6.7,	11.2,	1.7
66	101343.00	85.0,	93.5,	92.8	28.2,	17.6,	11.3	4.4,	7.7,	1.8
67	108094.00	87.4,	95.2,	91.9	26.8,	16.5,	11.7	5.1,	6.1,	1.2
38	60396.00	85.0,	91.4,	86.1	39.0,	26.6,	16.1	3.4,	4.6,	1.4

Table 2
Relative performances for the month of September

Domain size	Domain total(Rs.)	ACP I	ACP II	ACP III	$10^2 ACV$ I	$10^2 ACV$ II	$10^2 ACV$ III	RE (a)	RE (b)	RE (c)
30	24545.00	90.8,	94.2,	96.4	43.1,	45.5,	17.1	2.9,	13.5,	4.6
80	57622.00	78.2,	99.3,	100.0	25.0,	24.2,	11.4	9.4,	100.7,	10.7
156	124307.00	76.3,	98.9,	100.0	16.6,	13.6,	9.0	11.8,	45.5,	3.9
146	119680.00	79.4,	99.2,	100.0	17.7,	14.1,	9.1	12.8,	56.6,	4.4
107	99483.00	88.8,	99.5,	100.0	21.6,	16.3,	10.2	9.4,	42.9,	4.5
130	140299.00	90.8,	98.2,	99.0	20.0,	13.1,	9.2	6.7,	18.5,	2.7
106	127003.00	90.9,	98.8,	99.0	21.5,	14.5,	9.6	9.2,	18.6,	2.0
80	115944.00	87.0,	94.7,	94.9	26.9,	16.1,	10.4	5.0,	9.1,	1.8
63	95144.00	86.2,	94.1,	93.0	28.0,	18.4,	11.1	4.0,	7.2,	1.8
52	79870.00	82.6,	91.1,	87.0	34.2,	21.7,	13.5	3.7,	5.3,	1.4
54	99590.00	82.4,	90.9,	87.8	36.5,	18.7,	11.6	4.5,	7.1,	1.6

4. Comments and Recommendations

If three criteria are acceptable then an obvious conclusion from the two tables given in section 3 is that, though the models are postulated drastically without even trying to check their validity, one should employ empirical Bayes estimators starting with initial designed-based estimators expecting appreciable improvement and also employ Kalman filtering for a further improvement in situations treated as above.

5. Appendix

A NOTE ON RAO, HARTLEY, COCHRAN (RHC) SAMPLING STRATEGY AND ITS MODIFICATION IN TWO STAGE SAMPLING

If a sample of size n is to be taken from a population of size N to estimate the total Y of values y_i, $(1, \cdots, N)$ of a variable y when 'normed' size-measures p_i, $\left(0 < p_i < 1, \sum_1^N p_i = 1, i = 1, \cdots, N\right)$ are available, then RHC's method of sampling is as follows. The population is divided at random into n groups of sizes N_g $(g = 1, \cdots, n; 0 < N_g < N, \sum_1^n N_g = N)$.

From each group one unit is then chosen with a probability proportional to
values of p_i for the units falling within the group; selection is independent
across the groups. Writing y_g, p_g for y_i and p_i-value of the unit chosen from
the g- th group and c_g for the sum of the p_i - values over the units falling in
the g -th group and $\sum_g$ for the sum over the n groups, the RHC estimator
for Y is

$$t_R = \sum_g y_g \frac{c_g}{p_g} .$$

An unbiased estimator of the variance of t_R is

$$v(t_R) = \left(\frac{\sum_g N_g^2 - N}{N^2 - \sum_g N_g^2} \right) \sum_g c_g \left(\frac{y_g}{p_g} - t_R \right)^2 .$$

To control the variance of t_R an appropriate choice of N_g for each g is $\frac{N}{n}$
if it is an integer; otherwise take some of them as $\left[\frac{N}{n} \right]$ and rest as $\left[\frac{N}{n} \right] + 1$
subject to $\sum_g N_g = N$.

If the unit i consists of M_i second stage units (ssu), one instead of ascer-
taining y_i may estimate it by taking an SRSWOR of size m_i $(0 < m_i < M_i,$
$i = 1, \cdots, N)$, employing the expansion estimator $\hat{y}_i$. Writing s_{2i}^2 for the
sample variance, using divisor $(m_i - 1)$, an estimator of variance of $\hat{y}_i$ based
on selected ssu's is

$$\hat{v}(\hat{y}_i) = M_i (M_i - m_i) \frac{s_{2i}^2}{m_i} .$$

In such a case one may take

$$\hat{t}_R = \sum_g \hat{y}_g \frac{c_g}{p_g}$$

instead of t_R as an estimator for Y. Then a standard unbiased estimator of
the variance of $\hat{t}_R$ with obvious notations is

$$\hat{v}(\hat{t}_R) = \left(\frac{\sum_g N_g^2 - N}{N^2 - \sum_g N_g^2} \right) \sum_g c_g \left(\frac{\hat{y}_g}{p_g} - \hat{t}_R \right)^2 + \sum_g c_g \frac{\hat{v}(\hat{y}_g)}{p_g} .$$

For our application as reported in sections 2 and 3 this procedure is followed
separately within each statum treated as a population and y_i is replaced
throughout by y_{di} and thus formulae for t_d and v_d of sections 2 and 3 are
derived.

BIBLIOGRAPHY

Ghosh, M. and Rao, J.N.K. 1994). Small area estimation: an appraisal.
     Statistical Sc. **81**, 1058-1062.

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

Prasad, N.G.N. and Rao, J.N.K. (1990). The estimation of mean squared errors of small area estimators. *Jour. Amer. Stat. Assoc.* **85**, 163-171.

Rao, J.N.K., Hartley, H.O. and Cochran, W.G. (1962). On a simple procedure of unequal probability sampling without replacement. *Jour. Roy. Stat. Soc.* B, **24**, 482-431.