Journal of Official Statistics, Vol. 19, No. 3, 2003, pp. 253–263

Simultaneous Estimation of the Mean of a Binary Variable from a Large Number of Small Areas

Li-Chun Zhang[1]

We develop a frequentist method of simultaneous small area estimation under hierarchical models. The simultaneous estimator is the best ensemble predictor under the model. It is preferable to the area-specific best predictor when the distribution of the small area parameters is of interest as well. We provide details of application to binary data. We illustrate the proposed methodology on register employment and unemployment data, and validate the results by the true population values.

Key words: Ensemble estimator; hierarchical model; bootstrap.

1. Introduction

Small area estimation methods are often developed from the point of view of area-specific prediction. (See Ghosh and Rao (1994) for an excellent overview.) Such methods, however, may be unsatisfactory if we are interested in the distribution of the small area parameters as well. For instance, the between-area variation of the estimates is often found to be much smaller than the true variation in the population, which is known as over-shrinkage. Previously, the problem has been dealt with under the Bayesian framework (Louis 1984; Spjøtvoll and Thomsen 1987; Lahiri 1990; Ghosh 1992). Judkins and Liu (2000) carried out a simulation study for binary data, focusing on the estimator of Spjøtvoll and Thomsen (1987) and a simple version of the constrained empirical Bayes (CEB) estimator of Lahiri (1990), both of which were found to give better estimates of the range of the small area parameters compared to the direct estimator and the empirical Bayes (EB) estimator.

In this article we develop a method of simultaneous estimation from the frequentist perspective, which recognizes the finiteness of the population that we are dealing with. More explicitly, let θ_i be the parameter of interest from area i, for $i = 1, \ldots, A$. The collection of all the θ_i's, denoted by $\{\theta_i\}$, forms an ensemble (Judkins and Liu 2000). Now that the ensemble is of a finite size, one of them must have the smallest value among all the θ_i's, another must have the second smallest value, and so on. This is a fact that we may condition on in small area estimation. Since the rank of a small area parameter is defined together with the ranks of all the other small area parameters, the best ensemble

[1] Statistics Norway, Kongens gt. 6, P.B. 8131 Dep, N-0033 Oslo, Norway. Email: lcz@ssb.no
Acknowledgment: I am very grateful to an anonymous referee and the associate editor for helpful suggestions, which greatly improved the presentation.

predictor of $\{\theta_i\}$ must be derived simultaneously. Hence the term simultaneous estimation. We explain the details of the proposed methodology in Section 3.

But first, we review some area-specific prediction methods in Section 2. We start with the model of Fay and Herriot (1979). We show that the best linear unbiased predictor (BLUP) entails loss of between-area variation, just like the Bayes estimator (Louis 1984). Moreover, the Fay-Herriot model resembles the EB approach in that we assume only the first two moments of θ_i, without specifying its distribution. This is inadequate if the interested distribution of θ_i is nonnormal. We turn next to two-level hierarchical models. At the lower level, we have a distribution of θ_i; at the upper level, we have a conditional distribution of the data given θ_i. We review the beta-binomial model for binary data, which will be used in the numerical study in Section 4. While it removes some of the difficulties of the Fay-Herriot model, we show that the area-specific best predictor under the beta-binomial model also suffers from over-shrinkage.

Finally, we apply the method of simultaneous estimation to the register employment and unemployment data of Norway. We use the register variables instead of the Labor Force Survey (LFS) variables for this study because it allows us to validate the results by the true values in the population. Comparisons will be made with the direct estimator, the frequentist area-specific predictors under both the Fay-Herriot model and the beta-binomial model, the EB estimator, and the Lahiri CEB estimator.

2. Some Area-specific Prediction Methods

2.1. *Best linear unbiased predictor*

Denote by θ_i the mean of a binary variable from small area i, for $i = 1, \ldots, A$. Let $\tilde{\theta}_i$ be a direct estimator such that $E_D(\tilde{\theta}_i \mid \theta_i) = \theta_i$, where E_D denotes expectation with respect to the sampling design. Direct Fay-Herriot modeling gives us

$$\theta_i = \mu + \nu_i \qquad \text{and} \qquad \tilde{\theta}_i = \theta_i + e_i$$

where ν_i is the area-level random effect with zero mean and variance σ^2, and e_i is the sampling error with zero mean and variance $\psi_i = \mathrm{Var}_D(e_i \mid \theta_i)$. When both ψ_i and σ^2 are known, the best linear unbiased predictor (BLUP) is given by

$$\hat{\theta}_i = (1 - \gamma_i)\hat{\mu} + \gamma_i \tilde{\theta}_i \qquad \text{where } \gamma_i = \sigma^2/(\psi_i + \sigma^2)$$

and $\hat{\mu}$ is the generalized least square (GLS) estimate of μ. In practice, σ^2 is usually unknown. Let $\hat{\sigma}^2$ be a suitable estimator of σ^2. The plug-in BLUP, where σ^2, is replaced by $\hat{\sigma}^2$, is referred to as the empirical best linear unbiased predictor (EBLUP).

Both the direct estimator and the BLUP are derived for area-specific prediction. Neither of them leads to the right amount of between-area variation. This is easy to show in the case of $\psi_i = \psi$, where we have $\gamma_i = \gamma$ and

$$\mathrm{Var}_{M,D}(\tilde{\theta}_i) = E_M[\mathrm{Var}_D(\tilde{\theta}_i \mid \theta_i)] + \mathrm{Var}_M[E_D(\tilde{\theta}_i \mid \theta_i)] = \psi + \sigma^2 = \sigma^2/\gamma$$

where E_M and Var_M denote expectation and variance with respect to the model. Thus,

$$E_{M,D}\left[(A-1)^{-1}\sum_{i=1}^{A}(\tilde{\theta}_i - \bar{\tilde{\theta}})^2\right] = \frac{A}{A-1}\mathrm{Var}_{M,D}(\tilde{\theta}_i - \bar{\tilde{\theta}}) = \mathrm{Var}_{M,D}(\tilde{\theta}_i) = \sigma^2/\gamma$$

where $\tilde{\bar{\theta}} = \sum_i \tilde{\theta}_i/A$ is the mean of the direct estimators, whereas

$$E_{M,D}\left[(A-1)^{-1}\sum_{i=1}^{A}(\hat{\theta}_i - \hat{\bar{\theta}}_i)^2\right] = \gamma^2 \frac{A}{A-1}\mathrm{Var}_{M,D}(\tilde{\theta}_i - \tilde{\bar{\theta}}_i) = \gamma^2\,\mathrm{Var}_{M,D}(\tilde{\theta}_i) = \gamma\sigma^2$$

where $\hat{\bar{\theta}} = \sum_i \hat{\theta}_i/A$ is the mean of the BLUPs. In other words, while the direct estimators are over-dispersed by a factor of $1/\gamma$, the BLUPs are under-dispersed by a factor of γ. The results are similar to those under the Bayesian framework (Louis 1984).

Some additional difficulties of the EBLUP approach are worth noticing. Firstly, only the first two moments of θ_i are specified in the Fay-Herriot model. When we are interested in estimating the distribution of θ_i, such a model is inadequate if the distribution of θ_i is quite different from the normal distribution. This is for instance the case when θ_i is the small area unemployment rate, which is close to 0 and has a skewed distribution.

Secondly, not only σ^2 but also ψ_i is usually unknown and needs to be estimated. Denote by n_i the sample size in area i. For binary data, let $y_{ij} \in \{0, 1\}$ be the value of the jth unit, for $j = 1, \ldots, n_i$. Let $y_{i.} = \sum_{j=1}^{n_i} y_{ij}$ and $\bar{y}_{i.} = y_{i.}/n_i$. Under some noninformative sampling design, small n_i implies a large probability of obtaining a *degenerate* subsample where $y_{i.} = 0$ or n_i, especially if θ_i is close to 0 or 1. In such cases, the direct estimate of ψ_i is 0, so the EBLUP is the same as $\tilde{\theta}_i$, which can be very misleading. While this can be avoided by using some biased, *ad hoc* smooth estimator of ψ_i, no theoretically sound solution within the EBLUP approach seems available at the moment.

2.2. *Best predictor under beta-binomial model*

Some of the difficulties of the EBLUP can be removed by means of hierarchical modeling. For binary data, we assume that $Y_{i.}$ has the Binomial(n_i, θ_i) distribution. There are many possibilities for the distribution of θ_i. The logit-normal distribution and the beta distribution are two of the most commonly used. For the theoretical discussion below and the numerical study later in this article, we will use the beta-binomial model because it is analytically more tractable. However, we notice that the method of simultaneous estimation that we develop is applicable to any hierarchical model of choice.

The beta-binomial model is given as

$$\theta_i \sim \mathrm{Beta}(\nu, \omega) \qquad \text{and} \qquad Y_{i.} \mid \theta_i \sim \mathrm{Binomial}(n_i, \theta_i)$$

The direct estimator is $\tilde{\theta}_i = \bar{y}_{i.}$ with sampling variance $\psi_i = \theta_i(1-\theta_i)/n_i$. If the sample division of $(n_i, y_{i.})$ is sufficient so that, within each area, we have approximately simple random sampling, then the beta-binomial model can be considered to contain both model and design variation. Otherwise, it is strictly a modeling approach.

The joint distribution of $(y_{i.}, \theta_i)$ is given as

$$f(y_{i.}, \theta_i) = \left\{\binom{n_i}{y_{i.}}\theta_i^{y_{i.}}(1-\theta_i)^{n_i - y_{i.}}\right\}\{\theta_i^{\nu-1}(1-\theta_i)^{\omega-1}/B(\nu,\omega)\}$$

where $B(\nu, \omega)$ is the beta function. We obtain the marginal distribution of $y_{i.}$ by integrating out θ_i. The log-likelihood, denoted by $l(\nu, \omega)$ is then given by

$$\begin{aligned} l(\nu,\omega) &= \sum_i \log B(y_{i.}+\nu, n_i - y_{i.}+\omega) - A\cdot \log B(\nu,\omega) \\ &= \sum_i \left\{ \log\frac{\Gamma(y_{i.}+\nu)}{\Gamma(\nu)} + \log\frac{\Gamma(n_i-y_{i.}+\omega)}{\Gamma(\omega)} - \log\frac{\Gamma(n_i+\nu+\omega)}{\Gamma(\nu+\omega)} \right\} \\ &= \sum_i \left\{ \sum_{h=0}^{y_{i.}-1} \log(h+\nu) + \sum_{h=0}^{n_i-y_{i.}-1} \log(h+\omega) - \sum_{h=0}^{n_i-1} \log(h+\nu+\omega) \right\} \end{aligned}$$

where $\Gamma(\cdot)$ is the gamma function, and any sum with an upper limit of -1 is set to be zero. We obtain the maximum likelihood estimate (MLE), i.e., $(\hat{\nu}, \hat{\omega})$, by the Newton-Raphson method, and its covariance matrix by the inverse of the observed information matrix.

The best predictor of θ_i under the beta-binomial model is the conditional expectation of θ_i given $y_{i.}$. The beta distribution being the conjugate prior of the binomial distribution, the conditional distribution of θ_i given $y_{i.}$ is again a beta distribution. We have

$$\theta_i \mid y_{i.} \sim \text{Beta}(y_{i.}+\nu, n_i - y_{i.}+\omega) \qquad \text{and} \qquad p_i = E(\theta_i \mid y_{i.}) = \frac{y_{i.}+\nu}{n_i+\nu+\omega}$$

Thus, skewness of θ_i and degenerate subsamples $(n_i, y_{i.})$ are handled at the same time. The best predictor can also be written as a weighted sum between the overall mean and the direct estimator. For simplicity, denote by E and Var the model expectation and variance. Let $\mu = E(\theta_i) = \nu/(\nu+\omega)$ and $\tau = 1/(\nu+\omega)$. We have

$$p_i = (1-\gamma_i)\mu + \gamma_i \bar{y}_{i.} \qquad \text{where } \gamma_i = n_i/(1/\tau + n_i)$$

Since $\sigma^2 = \text{Var}(\theta_i) = \mu(1-\mu)\tau/(\tau+1)$, we have, for large $\nu+\omega$

$$\tau/(\tau+1) = 1/(1+\nu+\omega) \approx 1/(\nu+\omega) = \tau \Rightarrow \sigma^2 \approx \mu(1-\mu)\tau$$

and by substitution

$$\gamma_i \approx \sigma^2/(\phi_i + \sigma^2) \qquad \text{where } \phi_i = \mu(1-\mu)/n_i$$

It is now easy to show that, if $n_i = n$, then $\phi_i = \phi$, and $\gamma_i = \gamma$, and

$$E\left[(A-1)^{-1}\sum_{i=1}^{A}(\tilde{\theta}_i - \bar{\tilde{\theta}})^2\right] = \frac{A}{A-1}\text{Var}(\tilde{\theta}_i - \bar{\tilde{\theta}}) = \text{Var}(\tilde{\theta}_i) = E(\psi_i) + \text{Var}(\theta_i) \approx \sigma^2/\gamma$$

where $\bar{\tilde{\theta}} = \sum_i \tilde{\theta}_i/A$ is the mean of the direct estimators, whereas

$$E\left[(A-1)^{-1}\sum_{i=1}^{A}(p_i - \bar{p})^2\right] = \gamma^2\frac{A}{A-1}\text{Var}(\tilde{\theta}_i - \bar{\tilde{\theta}}) = \gamma^2\,\text{Var}(\tilde{\theta}_i) \approx \gamma\sigma^2$$

where $\bar{p} = \sum_i p_i/A$ is the mean of the best predictors. In other words, similar to the results under the Fay-Herriot model, the direct estimators are over-dispersed by a factor of $1/\gamma$, whereas the best predictors are under-dispersed by a factor of γ.

3. Simultaneous Small Area Estimation

3.1. Derivation of estimates

Simultaneous estimation consists of two steps. First, the best ensemble predictor is derived regardless of the specific areas. Next, instead of using the area-specific best predictors themselves, these are ranked, and the estimates of the ranks are used to match the best ensemble predictor with the small areas. In this way, the simultaneous estimates recover the amount of under-dispersion in the area-specific predictors.

Denote by $F(x_i, \theta_i; \xi)$ some hierarchical model, where x_i denotes the data from area i and ξ contains the parameters of the model. Let $F(\theta_i; \xi)$ be the marginal distribution of θ_i. Let $\hat{\xi}$ be the MLE of ξ. Let $\theta_{(i)}$ be the ith order statistic of $\{\theta_i\}$, where $\theta_{(1)} \leqslant \theta_{(2)} \leqslant \cdots \leqslant \theta_{(A)}$. Let the expectation of $\theta_{(i)}$ and its estimator be, respectively,

$$\eta_i = E(\theta_{(i)}; \xi) \qquad \text{and} \qquad \hat{\eta}_i = E(\theta_{(i)}; \hat{\xi}) \tag{1}$$

Since η_1 is the best predictor of $\theta_{(1)}$, and η_2 is the best predictor of $\theta_{(2)}$, and so on, $\{\eta_i\}$ is the best ensemble predictor of $\{\theta_i\}$ under the distribution $F(\theta_i; \xi)$, and $\{\hat{\eta}_i\}$ is the estimated best ensemble predictor. We notice that, when the number of small areas is large, i.e., as $A \to \infty$, η_i is approximately given by

$$\eta_i = E(\theta_{(i)}; \xi) \approx F^{-1}(\alpha_i; \xi) \qquad \text{where } \alpha_i = i/(A+1)$$

(Balakrsihnan and Rao 1998), and $F^{-1}(\alpha; \xi)$ is the α-quantile of $F(\theta_i; \xi)$.

For each $i = 1, \ldots, A$, η_i is derived with respect to the conditional distribution

$$F(\theta_j; \xi \mid rank(\theta_j) = i) = F(\theta_{(i)}; \xi)$$

It is not area-specific if we do not know which area is associated with $\theta_{(i)}$, which is the case in most applications. Since the rank of a small area parameter is defined together with the parameters of all the other areas, the best ensemble predictor $\{\eta_i\}$ is derived simultaneously. In contrast, the area-specific best predictor of θ_i is derived with respect to another conditional distribution, i.e., the distribution of θ_i given x_i, denoted by $F(\theta_i; \xi \mid x_i)$. The area-specific best predictors are thus derived independently of each other. In the extreme case where there are no data at all, the area-specific best predictor, given as $\mu = E(\theta_i; \xi)$, is identical for all the small areas, which is completely misleading as an ensemble estimator of $\{\theta_i\}$.

It remains to match $\{\hat{\eta}_i\}$ with the small areas. This is straightforward if the ranks of the small areas are known. Otherwise, let $\hat{\theta}_i = E(\theta_i; \hat{\xi} \mid x_i)$ be the estimated area-specific predictor of θ_i. We use the rank of $\hat{\theta}_i$ among $\{\hat{\theta}_i\}$ as an estimate of the rank of θ_i among $\{\theta_i\}$. The *simultaneous estimate* of θ_i, denoted by $\dot{\theta}_i$, is then given by

$$\dot{\theta}_i = \hat{\eta}_{r_i} \qquad \text{where } r_i = rank(\hat{\theta}_i) \tag{2}$$

The simultaneous estimates have therefore the same order as the estimated area-specific predictors. In the special case of ties among the estimated area-specific predictors, we assign the corresponding $\hat{\eta}_i$'s by random permutation.

Consider now the properties of $\dot{\theta}_i$ as an area-specific predictor. Asymptotically, let the sample sizes tend to infinity in all the areas. Assume that the MLE $\hat{\xi}$ is a consistent estimator of ξ, and the area-specific predictor $\hat{\theta}_i$ is consistent for θ_i. It follows that $rank(\hat{\theta}_i)$ is a consistent estimator of the true rank of θ_i, denoted by $d_i = rank(\theta_i)$, and the simultaneous estimator $\dot{\theta}_i$ is a consistent estimator of η_{d_i}. While η_{d_i} is not equal to θ_i, the difference between the two tends to zero in probability as $A \to \infty$, because the variance of $\theta_{(d_i)}$ tends to zero in this case (Balakrishnan and Rao 1998).

3.2. Derivation of predictive intervals

Given the data x_i, we may use the quantiles of $F(\theta_i; \hat{\xi} \mid x_i)$ to estimate the quantiles of $F(\theta_i; \xi \mid x_i)$, and then derive the predictive intervals of θ_i conditional on x_i. It is possible to add an adjustment similar to that in (2).

Consider the following finite population bootstrap:

1. obtain $\{\theta_i^*\}$ where $\theta_i^* \sim F(\theta_i; \hat{\xi} \mid x_i)$ is independently generated for $i = 1, \ldots, A$;
2. set $\{\theta_i^{(1)}\}$ where $\theta_i^{(1)} = \hat{\eta}_{r_i}$ and r_i is the rank of θ_i^* among $\{\theta_i^*\}$ and $\hat{\eta}_i$ by (1).

Let M be the number of independent repetitions of Steps 1–2. We use the sample α-quantile of $\{\theta_i^{(1)}, \ldots, \theta_i^{(M)}\}$ as the bootstrap estimate of the α-quantile of θ_i.

Step 2 of the bootstrap above adjusts the ensemble property of the bootstrap replicates $\{\theta_1^{(k)}, \ldots, \theta_A^{(k)}\}$ for $k = 1, \ldots, M$, forcing their order statistics to be equal to $\{\hat{\eta}_i\}$. Its effect on the area-specific predictive intervals is unclear, however. Consider again the extreme case where there are no data at all but the distribution $F(\theta_i; \xi)$ is known. Without the adjustment, all the areas have the same bootstrap distribution $F(\theta_i; \xi)$. With the adjustment, a bootstrap replicate of θ_i takes the value η_j with the probability $P[rank(\theta_i) = j; \xi]$, which is simply a discrete version of $F(\theta_i; \xi)$. The predictive intervals will not differ much one way or the other if the number of small areas is large. In a given situation, therefore, we should derive the predictive intervals both with and without the finite population adjustment, and compare the results.

It is worth noticing that the predictive interval of θ_i typically is asymmetric around the simultaneous estimate $\dot{\theta}_i$. For areas with high estimated ranks, the interval is longer on the lower side of $\dot{\theta}_i$, whereas for areas with low estimated ranks, the interval is longer on the upper side of $\dot{\theta}_i$. Take for instance the area with the highest estimated rank, say, $\dot{\theta}_i = \hat{\eta}_A$. Since there are only A small areas in the population, the actual rank of θ_i can possibly be lower than A, but not higher. Given that the largest small area parameter in the population is about $\hat{\eta}_A$, the predictive interval of θ_i must be much longer on the lower side of $\dot{\theta}_i$. Again, this is a fact determined by the finiteness of the population.

4. Estimation of Municipality Register Employment and Unemployment Rates

4.1. Data and model

We consider the estimation of municipality employment and unemployment rates, based on the administrative register data in Norway. We obtain the sample by linking the registers to the LFS sample at the individual level, which is possible in several countries including Norway. For simplicity, we define the employment rate as the ratio between the employment and population totals, and likewise for the unemployment rate. In this study we have used the Norwegian LFS sample of the 4th Quarter in 1997 (with 21,676 individuals). Only 3 out of the total 435 municipalities are not represented in the sample and these will be left out of the analysis, so that $A = 432$ both in the sample and in the population.

There is a problem of nonsampling errors, when we compare the register population rates with the various estimates. For instance, nonresponse is known to depend on the LFS employment and unemployment status (Thomsen and Zhang 2001). To concentrate

Table 1. Quantiles of the municipality register employment and unemployment rates: (I) the population values, and (II) the beta distribution fitted to the population values

	Municipality register employment rate									
Quantile α	0.025	0.05	0.10	0.16	0.25	0.75	0.84	0.90	0.9	0.975
(I)	0.636	0.649	0.663	0.677	0.692	0.751	0.769	0.782	0.800	0.813
(II)	0.628	0.644	0.661	0.676	0.691	0.753	0.767	0.779	0.793	0.805
	Municipality register unemployment rate									
Quantile α	0.025	0.05	0.10	0.16	0.25	0.75	0.84	0.90	0.95	0.975
(I)	0.007	0.008	0.010	0.012	0.015	0.026	0.029	0.032	0.039	0.047
(II)	0.007	0.008	0.010	0.012	0.014	0.027	0.030	0.034	0.039	0.043

on the issue at hand, we adjust the population rates as follows. Let θ_i^R be the true rate of interest in the ith municipality. Let $(n_i, y_{i.})$ be the corresponding sample. We put

$$\theta_i = \theta_i^R - \sum_i \theta_i^R \Big/ A + \left(\sum_i y_{i.}\right) \Big/ \left(\sum_i n_i\right)$$

While this removes most of the bias due to the nonsampling errors, it does not affect the finite population distribution of interest except for a shift in location. Notice that we do not adjust the sample values because of problems with the degenerate subsamples.

As an exploratory analysis we fit the beta distribution to $\{\theta_i\}$. Table 1 gives the true quantiles of θ_i compared to the ones derived from the fitted beta distributions. For the employment rates, the beta distribution is slightly long-tailed at the lower end and slightly short-tailed at the upper end, whereas for the unemployment rates the beta distribution fits the actual population almost exactly. We conclude that the beta distribution is a reasonable model for the population rates which we are considering.

4.2. *Estimates*

We compare the simultaneous estimator under the beta-binomial model to the direct estimator, the EBLUP under the Fay-Herriot model, the EB estimator, the Lahiri CEB estimator, and the estimated area-specific best predictor under the beta-binomial model. We use the *averaged squared error (ASE)* to measure how an estimator works for area-specific prediction, and we use the *averaged squared distributional error (ASDE)* to measure how an estimator performs as an ensemble estimator. For the direct estimator, we have

$$ASE(\tilde{\theta}_i) = A^{-1} \sum_{i=1}^{A} (\tilde{\theta}_i - \theta_i)^2 \qquad \text{and} \qquad ASDE(\tilde{\theta}_i) = A^{-1} \sum_{i=1}^{A} (\tilde{\theta}_{(i)} - \theta_{(i)})^2$$

where $\tilde{\theta}_{(i)}$ is the ith order statistic of $\{\tilde{\theta}_{(i)}\}$ and similarly for the other estimators.

The root ASEs and ASDEs of all the estimators are given in Table 2. The direct estimator is clearly the worst method. The EBLUP is so sensitive towards degenerate subsamples that it gives very misleading results both for employment and unemployment rates. The area-specific best predictor relates to the simultaneous estimator in very much the same way as the EB estimator relates to the Lahiri CEB estimator. The area-specific best

Table 2. Root ASE and ASDE of the estimated municipality register employment and unemployment rates: (A) the direct estimator, (B) the EBLUP, (C) the EB estimator, (D) the Lahiri CEB estimator, (E) the area-specific best predictor under the beta-binomial model, and (F) the simultaneous estimator under the beta-binomial model

	Municipality register employment rate					
Estimator	A	B	C	D	E	F
Root ASE	0.139	0.105	0.042	0.051	0.042	0.050
Root ASDE	0.104	0.071	0.021	0.006	0.022	0.005
	Municipality register unemployment rate					
Estimator	A	B	C	D	E	F
Root ASE	0.0340	0.0191	0.0095	0.0094	0.0093	0.0098
Root ASDE	0.0269	0.0170	0.0090	0.0065	0.0081	0.0039

predictor and the EB estimator are only better for area-specific prediction of register employment rates. The simultaneous estimator and the Lahiri CEB estimator are much better ensemble estimators for both employment and unemployment rates. Finally, the simultaneous estimator is better than the Lahiri CEB estimator for register unemployment rates. Notice that we have used a simple version of the Lahiri CEB estimator, which was implemented for binary data by Judkins and Liu (2000). Although it does not exactly satisfy the ''posterior linearity'' condition for the Lahiri CEB estimator, it worked fairly well in the simulation study of Judkins and Liu (2000).

Table 3 gives the quantiles of the estimated municipality register employment and unemployment rates, which may be compared to the true quantiles given in Table 2. As

Table 3. Quantiles of the estimated municipality register employment and unemployment rates: (A) the direct estimator, (B) the EBLUP, (C) the EB estimator, (D) the Lahiri CEB estimator, (E) the area-specific best predictor under the beta-binomial model, and (F) the simultaneous estimator under the beta-binomial model

	Municipality register employment rate									
Quantile α	0.025	0.05	0.10	0.16	0.25	0.75	0.84	0.90	0.95	0.975
A	0.375	0.500	0.556	0.600	0.641	0.800	0.833	0.864	1.000	1.000
B	0.613	0.642	0.658	0.676	0.689	0.773	0.794	0.823	1.000	1.000
C	0.669	0.678	0.688	0.696	0.702	0.734	0.743	0.749	0.759	0.766
D	0.621	0.640	0.658	0.675	0.686	0.748	0.767	0.779	0.797	0.812
E	0.670	0.681	0.690	0.698	0.703	0.733	0.742	0.748	0.757	0.764
F	0.623	0.639	0.657	0.672	0.688	0.751	0.766	0.778	0.792	0.803
	Municipality register unemployment rate									
Quantile α	0.025	0.05	0.10	0.16	0.25	0.75	0.84	0.90	0.95	0.975
A	0.000	0.000	0.000	0.000	0.000	0.027	0.039	0.056	0.083	0.100
B	0.000	0.000	0.000	0.000	0.000	0.010	0.011	0.014	0.016	0.020
C	0.019	0.020	0.020	0.020	0.020	0.021	0.021	0.021	0.021	0.022
D	0.015	0.016	0.018	0.018	0.019	0.022	0.023	0.024	0.026	0.028
E	0.017	0.018	0.019	0.019	0.019	0.021	0.021	0.022	0.023	0.024
F	0.010	0.011	0.013	0.014	0.016	0.024	0.026	0.028	0.031	0.033

was shown in Section 2, the direct estimator leads to over-dispersion, whereas the area-specific best predictors lead to under-dispersion. The strange behavior of the EBLUP is due to its sensitivity towards degenerate subsamples, where the EBLUPs are misleadingly set to be the direct estimates. Both the Lahiri CEB estimator and the simultaneous estimator give good estimates of the quantiles of the register employment rates. However, both are seen to under-estimate the between-area variation of the register unemployment rates. Since the exploratory analysis shows that the beta-binomial model is appropriate for the unemployment rates, we believe that the problem is caused by the extremely large number of degenerate subsamples, i.e., in 260 out of the 432 municipalities. In comparison, there are only 30 municipalities with degenerate subsamples in the case of employment.

4.3. *Predictive intervals*

Under the beta-binomial model, we have $\theta_i \sim \text{Beta}(y_{i.} + \nu, n_i - y_{i.} + \omega)$ conditional on $(n_i, y_{i.})$. We first obtain the estimated quantiles of this beta distribution where (ν, ω) are replaced by the MLE. Next, we derive the quantiles using the finite population bootstrap method described in Section 3. Finally, we calculate the coverage levels of these quantiles. More explicitly, denote by $\hat{\theta}_i(\alpha)$ the α-quantile of θ_i under the estimated conditional beta distribution, we derive its coverage level as

$$A^{-1} \sum_i I_{\theta_i \leqslant \hat{\theta}_i(\alpha)} \qquad \text{where } I_{\theta_i \leqslant \hat{\theta}_i(\alpha)} = 1 \text{ if } \theta_i \leqslant \hat{\theta}_i(\alpha) \text{ and 0 otherwise.}$$

The approach for the finite population bootstrap α-quantiles is similar.

The coverage levels of the estimated quantiles are given in Table 4. There is little difference between the two methods. The finite population adjustment has no effect on the predictive intervals in the present case. The estimated quantiles attain almost the nominal coverage levels for the employment rates. The coverage levels are not satisfactory for the unemployment rates. This is due to the under-estimation of the between-area variation of the municipality unemployment rates, as reported in Table 3.

Finally, Figure 1 plots the true municipality register employment rates in the population, together with the estimated area-specific best predictors and the simultaneous estimates under the beta-binomial model, and the two-sided 90% predictive intervals. The

Table 4. Coverage level of the estimated quantiles of municipality register employment and unemployment rates: (I) the conditional beta distribution, (II) the finite population bootstrap (with 2200 bootstrap samples)

	Municipality register employment rate									
Quantile α	0.025	0.05	0.10	0.16	0.25	0.75	0.84	0.90	0.95	0.975
(I)	0.019	0.053	0.102	0.153	0.227	0.729	0.822	0.880	0.928	0.961
(II)	0.021	0.051	0.100	0.157	0.227	0.729	0.815	0.877	0.924	0.958
	Municipality register unemployment rate									
Quantile α	0.025	0.05	0.10	0.16	0.25	0.75	0.84	0.90	0.95	0.975
(I)	0.090	0.130	0.178	0.261	0.301	0.657	0.750	0.813	0.884	0.912
(II)	0.093	0.130	0.183	0.243	0.299	0.660	0.755	0.810	0.882	0.907

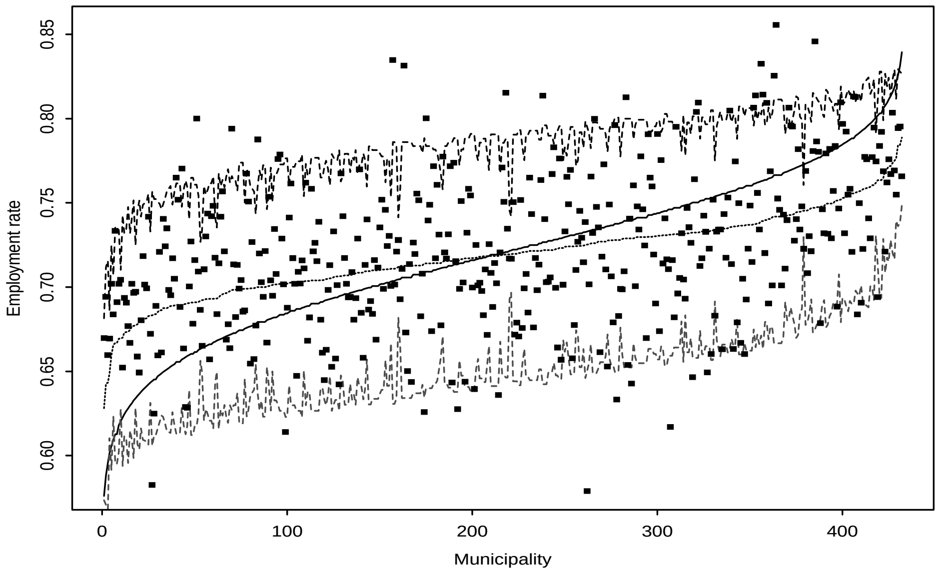

Figure 1. Two-sided 90% predictive intervals (dashed) and the population municipality register employment rates (square points): sorted in increasing order of the simultaneous estimates (solid line) and the area-specific best predictors (dotted line) under the beta-binomial model

simultaneous estimator is clearly the better ensemble estimator in terms of the range, the dispersion, and the quantiles, etc. Notice that the predictive intervals are much more asymmetric about the simultaneous estimates than the area-specific predictors.

5. Summary

We showed, in theory as well as by numerical example, that the area-specific best predictor entails loss of between-area variation of the small area parameters. We derived the simultaneous estimator as the best ensemble predictor under the assumed model. We illustrated the proposed method on binary data, and validated the results using the true population values. The numerical study has revealed a problem, where a large number of degenerate subsamples caused under-estimation of the between-area variation of the unemployment rates, even though the beta-binomial model was appropriate in that case. Given that the sample size can be quite small in at least some of the areas, degenerate subsamples seem unavoidable for categorical data, especially if the probabilities of some of the categories are close to 0 or 1. It is therefore a problem, which needs to be studied in more detail in the future. In addition, we plan to generalize the simultaneous estimation to situations where the survey variable is associated with known covariates.

6. References

Balakrishanan, N. and Rao, C.R. (1998). Order Statistics: An Introduction. In Order Statistics: Theory & Methods by R. Balakrishnan and C.R. Rao (eds), 3–24. Amsterdam: North-Holland.

Fay, R.E. and Herriot, R.A. (1979). Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data. Journal of the American Statistical Association, 74, 269–277.

Ghosh, M. (1992). Constrained Bayes Estimation with Applications. Journal of the American Statistical Association, 87, 533–540.

Ghosh, M. and Rao, J.N.K. (1994). Small Area Estimation: An Appraisal (with Discussion). Statistical Science, 9, 55–93.

Judkins, D.R. and Liu, J. (2000). Correcting the Bias in the Range of a Statistic Across Small Areas. Journal of Official Statistics, 16, 1–13.

Lahiri, P. (1990). Adjusted Bayes and Empirical Bayes Estimation in Finite Population Sampling. Sankhyā B, 42, 50–66.

Louis, T. (1984). Estimating a Population of Parameter Values Using Bayes and Empirical Bayes Methods. Journal of the American Statistical Association, 79, 393–398.

Spjøtvoll, E. and Thomsen, I. (1987). Application of Some Empirical Bayes Methods to Small Area Statistics. Bulletin of the International Statistical Institute, 2, 435–449.

Thomsen, I. and Zhang, L.-C. (2001). The Effects of Using Administrative Registers in Economic Short Term Statistics: The Norwegian Labour Force Survey as a Case Study. Journal of Official Statistics, 17, 285–294.

Received December 2000
Revised May 2003