

SMALL AREA ESTIMATION WITH UNMATCHED SAMPLING AND LINKING MODELS

Yong You¹ and J.N.K. Rao²

ABSTRACT

In small area estimation, a sampling error model on transformed survey estimates (e.g., log-transformation) and a corresponding linking model on the transformed population means are often used. This leads to a linear mixed effects combined model on the transformed estimates, and standard results in empirical Bayes (EB) or hierarchical Bayes (HB) can be applied to get improved small area estimates. But the assumption of zero mean sampling errors may not be valid in small samples due to nonlinear transformation. To avoid this difficulty, we propose a sampling error model on the original survey estimates and then combine it with the linking model on the transformed population means using a complete HB approach. A numerical study is conducted to compare the posterior means and posterior variances under the complete HB approach with those obtained under the customary linear mixed model approach.

KEY WORDS: Area-level model; Gibbs sampling; Linking model; Log-transformation; Sampling model; Small area.

RÉSUMÉ

L'estimation régionale utilise fréquemment des modèles pour les erreurs d'échantillonnage basées sur des transformations des estimations (transformations logarithmiques, par exemple) ainsi que les modèles de "liaisons" correspondants de moyennes de population transformées. Cela nous amène à utiliser un modèle à effets mixtes combinés sur les transformations des estimations, et des résultats connus sur les méthodes empiriques de Bayes (EB) ou les méthodes hiérarchiques de Bayes (HB) peuvent être alors employés pour améliorer les estimations régionales. Nous supposons que la moyenne des erreurs d'échantillonnage est de zéro mais cette supposition risque de ne pas être valide en raison de la transformation non-linéaire. Pour éviter cette difficulté, on propose d'utiliser un modèle pour les erreurs d'échantillonnage des estimations originales que l'on combine avec le modèle de "liaison" des moyennes de population transformées en utilisant l'approche complète HB. Une étude numérique est proposée pour comparer les moyennes et les variances a posteriori sous l'approche complète HB avec celles obtenues en utilisant le modèle linéaire mixte usuel.

MOTS CLÉS : Modèle au niveau de la région; échantillonnage de Gibbs; modèle de "liaison"; transformation logarithmique; modèle d'échantillonnage; petite région.

1. INTRODUCTION

Sample surveys are used to provide estimates not only for the total population but also for a variety of sub-populations (domains). Direct survey estimators, based only on the domain-specific sample data, are typically used to estimate parameters for large domains. But sample sizes in small domains, particularly small geographical areas, are rarely large enough to provide reliable direct estimates for specific small domains. In making estimates for small areas, it is necessary to "borrow strength" from related areas to form indirect estimators that increase the effective sample size and

thus increase the precision. Such indirect estimators are based on either implicit or explicit models that provide a link to related small areas through supplementary data such as recent census counts and current administrative records. It is now generally accepted that when indirect estimates are to be used they should be based on explicit models that relate the small areas of interest through supplementary data. Small area models may be broadly classified into two types: area-level and unit-level models. Ghosh and Rao (1994) and Rao (1999) presented a comprehensive overview and appraisal of models and methods for

¹ Yong You, Household Survey Methods Division, Statistics Canada, Ottawa, Canada, K1A 0T6, yongyou@statcan.ca.

² J.N.K. Rao, School of Mathematics and Statistics, Carleton University, Ottawa, Canada, K1S 5B6, jrao@math.carleton.ca.

small area estimation. In this paper, we focus on area-level models for small area estimation.

Suppose area-specific auxiliary data, x_i , are available for the sampled areas $i = 1, \dots, m$ as well as the nonsampled areas. A basic area-level model assumes that the population small area mean θ_i or some suitable function $\mu_i = g(\theta_i)$, such as $\mu_i = \log(\theta_i)$, is related to x_i through a linear model with random area effects v_i as follows:

$$\mu_i = x_i^T \beta + v_i, \quad i = 1, \dots, m, \quad (1)$$

where β is the p -vector of regression parameters and the v_i 's are uncorrelated with mean zero and variance σ_v^2 . Normality of the v_i is also often assumed. Model (1) is referred as a linking model for μ_i . It is also possible to partition the areas into groups and assume separate models of the form (1) across groups.

The basic area model also assumes that direct survey estimators y_i of small area mean θ_i are available whenever the area sample size $n_i > 1$. It is also customary to assume that

$$\hat{\mu}_i = \mu_i + e_i, \quad i = 1, \dots, m, \quad (2)$$

where $\hat{\mu}_i = g(y_i)$, and the sampling errors e_i are independent $N(0, \psi_i)$ with known sampling variance ψ_i .

Combining the sampling model (2) with the linking model (1), we get the well-known area level linear mixed model of Fay and Herriot (1979):

$$\hat{\mu}_i = x_i^T \beta + v_i + e_i, \quad i = 1, \dots, m. \quad (3)$$

Note that model (3) involves both design-based random variables e_i and model-based random variables v_i . Standard methods in empirical Bayes (EB) and hierarchical Bayes (HB) can be applied to (3) to obtain improved model-based small area estimates (Ghosh and Rao, 1994; Rao, 1999).

However, the assumption $E(e_i | \mu_i) = 0$ in the sampling model (2) may not be valid if the sample size n_i is small and μ_i is a nonlinear function of θ_i , even if the direct estimator y_i is design-unbiased for θ_i ,

i.e., $E(y_i | \theta_i) = \theta_i$. In the following sections, we consider a more realistic sampling model and study the effects of transformation through a HB approach.

2. AREA-LEVEL UNMATCHED MODELS

Let y_i denote the direct survey estimator of the i -th small area mean θ_i . We consider the following sampling model for y_i :

$$y_i = \theta_i + \varepsilon_i, \quad i = 1, \dots, m, \quad (4)$$

with $E(\varepsilon_i | \theta_i) = 0$, that is, the direct survey estimator y_i is design-unbiased for the small area mean θ_i . The sampling variance of y_i is $V(\varepsilon_i | \theta_i) = \theta_i^2 \psi_i^2$, where ψ_i^2 is known. Thus the coefficient of variation (CV) of y_i is ψ_i and does not depend on the small area mean θ_i . The standard deviation of y_i increases in direct proportion to the expected value θ_i .

The small area mean θ_i is assumed to be related to an area-level auxiliary variable x_i through a log-linear model with random area effects v_i as:

$$\log(\theta_i) = x_i^T \beta + v_i, \quad i = 1, \dots, m, \quad (5)$$

where β is a vector of unknown regression parameters, and the v_i 's are uncorrelated with $E(v_i) = 0$ and $V(v_i) = \sigma^2$, where σ^2 is unknown. Normality of v_i is also assumed. Fay and Herriot (1979) used a similar log-linear linking model in their application to estimating income for small areas. Note that we cannot directly combine the sampling model (4) with the linking model (5) to produce a linear mixed effects model for small area estimation as discussed in Section 1. As a result, standard results in linear model theory do not apply, unlike the case of model (3).

We now present the sampling model (4) and the linking model (5) in a hierarchical Bayes framework as follows:

$$y_i | \theta_i \sim N(\theta_i, \theta_i \psi_i^2), \quad i = 1, \dots, m; \quad (6)$$

and

$$\log(\theta_i) | \beta, \sigma^2 \sim N(x_i^T \beta, \sigma^2), \quad i = 1, \dots, m. \quad (7)$$

The linking model (7) implies that the small area mean θ_i conditionally has a log-normal distribution with density function given by

$$f(\theta_i | \beta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma\theta_i}} \exp\left\{-\frac{1}{2\sigma^2}(\log\theta_i - x_i^T \beta)^2\right\}. \quad (8)$$

We are interested in making inference on the small area means θ_i given the direct survey estimates y_i . For this purpose, we consider two approaches, namely, the customary transformation approach and a complete HB approach.

3. TRANSFORMATION APPROACH

In model (6), the sampling variance of y_i is proportional to θ_i^2 and is unknown since θ_i is unknown. Logarithmic transformation is usually used to stabilize the variance of y_i in which case the sampling variance is approximately equal to ψ_i if the area sample size is not small (Fay and Herriot, 1979). After the transformation, we can combine the linking model with the sampling model to produce a linear Fay-Herriot model. The resulting estimates based on the transformed data are then transformed back to the original scale. In general, variables are transformed using logarithms for two reasons: First, it is more plausible that the model is homoscedastic on the log scale (corresponding to a constant coefficient of variation or equal model variances of share in poverty) than on the original scale over the extremely wide range of county sizes. Second, the transformed variables tend to be symmetrically distributed, and the scatterplots of various covariates with the dependent variable are more linear. Recently, county-level (area-level) models with logarithmic transformation have been used to produce model-based county-level estimates of school-age children in poverty in the United States (National Research Council, 1997, 1998). Using these estimates, the US Department of Education allocates \$7 billion of federal funds to counties, and states distribute these funds among school districts in each county.

For the transformation approach, let $z_i = \log(y_i)$ and $\mu_i = \log(\theta_i)$. Then approximately we get $E(z_i) \approx \mu_i$ and $Var(z_i) \approx \psi_i^2$. Ignoring the approximation leads to the well-known Fay-Herriot model:

$$z_i | \mu_i \sim N(\mu_i, \psi_i^2), \quad i = 1, \dots, m; \quad (9)$$

and

$$\mu_i | \beta, \sigma^2 \sim N(x_i^T \beta, \sigma^2), \quad i = 1, \dots, m. \quad (10)$$

Models (9) and (10) are based on the transformed data z_i and parameter μ_i . Given μ_i , the transformed direct estimator z_i is assumed to be design-unbiased.

We are interested in estimating the posterior mean $E(\theta_i | Y)$ and the posterior variance $V(\theta_i | Y)$. Since $\theta_i = \exp(\mu_i)$, the first step is to obtain an estimator of μ_i based on the transformed models (9) and (10). The Gibbs sampling method is used to generate samples from the posterior distribution of μ_i . Suppose we have a sample $\{\mu_i^{(k)}, k = 1, \dots, G\}$ generated by the Gibbs sampler, then we have $\{\theta_i^{(k)}, k = 1, \dots, G\}$, with $\theta_i^{(k)} = \exp(\mu_i^{(k)})$ are samples from the posterior distribution of θ_i , and $E(\theta_i | Y)$ and $V(\theta_i | Y)$ can be estimated using the sample $\{\theta_i^{(k)}, k = 1, \dots, G\}$.

Assuming a flat prior for β and an inverse gamma $IG(a, b)$ for σ^2 , where a and b are known constants, we can obtain the following full conditional distributions required by the Gibbs sampler:

- $\mu | Z, \beta, \sigma^2 \sim N((1 - r_i)z_i + r_i x_i^T \beta, \psi_i^2 (1 - r_i))$,
where $r_i = \psi_i^2 / (\sigma^2 + \psi_i^2)$;
- $\beta | Z, \mu, \sigma^2 \sim N((\sum_{i=1}^m x_i x_i^T)^{-1} (\sum_{i=1}^m x_i \mu_i), \sigma^2 (\sum_{i=1}^m x_i x_i^T)^{-1})$;
- $\sigma^2 | Z, \mu, \beta \sim IG(a + m/2, b + \sum_{i=1}^m (\mu_i - x_i^T \beta)^2 / 2)$.

Since all the conditional distributions have closed form, implementation of the Gibbs sampler is straightforward. Thus by logarithmic transformation, the estimation procedure is simplified. However, bias could be introduced via transformation because the assumption may not be accurate for small sample sizes. In Section 4, we will consider a fully Bayesian approach on the original nonlinear model to evaluate this bias.

4. COMPLETE HIERARCHICAL BAYES APPROACH

Consider the original area-level models (6) and (7). We are interested in finding the posterior mean and posterior variance of θ_i given the data $Y = (y_1, \dots, y_m)^T$. From models (6) and (7), the joint density function of Y and parameters θ, β, σ^2 is given by

$$f(Y, \theta, \beta, \sigma^2) \propto \prod_{i=1}^m \frac{1}{\theta_i} \exp\left\{-\frac{(y_i - \theta_i)^2}{2\theta_i^2 \psi_i^2}\right\} \pi(\beta, \sigma^2) \cdot \prod_{i=1}^m \frac{1}{\sigma \theta_i} \exp\left\{-\frac{(\log \theta_i - x_i^T \beta)^2}{2\sigma^2}\right\} \quad (11)$$

where $\pi(\beta, \sigma^2)$ is the prior distribution of β and σ^2 as specified in Section 3. The joint distribution (11) can be employed to determine the following full conditional distributions required by the Gibbs sampler:

- $\theta_i | Y, \beta, \sigma^2 \propto \frac{1}{\theta_i^2} \exp\left\{-\frac{(y_i - \theta_i)^2}{2\theta_i^2 \psi_i^2} - \frac{(\log \theta_i - x_i^T \beta)^2}{2\sigma^2}\right\}$;
- $\beta | Y, \theta, \sigma^2 \sim N\left(\left(\sum_{i=1}^m x_i x_i^T\right)^{-1} \left(\sum_{i=1}^m x_i \log \theta_i\right), \sigma^2 \left(\sum_{i=1}^m x_i x_i^T\right)^{-1}\right)$;
- $\sigma^2 | Y, \theta, \beta \sim IG\left(a + m/2, b + \sum_{i=1}^m (\log \theta_i - x_i^T \beta)^2 / 2\right)$.

Unlike the case of Fay-Herriot model under transformation, the conditional distribution $[\theta_i | Y, \beta, \sigma^2]$ does not have a closed form. To draw samples from $[\theta_i | Y, \beta, \sigma^2]$, the Metropolis-Hastings updating scheme (see, e.g., Chib and Greenberg, 1995) is used within the Gibbs sampler. We note that the conditional density function of $[\theta_i | Y, \beta, \sigma^2]$ can be written as

$$\pi(\theta_i | Y, \beta, \sigma^2) \propto h(\theta_i) f(\theta_i | \beta, \sigma^2), \quad (12)$$

where $f(\theta_i | \beta, \sigma^2)$ is the log-normal density function given by (8) and $h(\theta_i)$ is a function given by

$$h(\theta_i) = \frac{1}{\theta_i} \exp\left\{-\frac{(y_i - \theta_i)^2}{2\theta_i^2 \psi_i^2}\right\}. \quad (13)$$

To implement the Metropolis-Hastings sampling scheme within the Gibbs sampler, we use $f(\theta_i | \beta, \sigma^2)$ as the candidate generating density function. Then at the k -th iteration, the Metropolis-Hastings probability of moving to $\theta_i^{(k+1)}$ is given by

$$\alpha(\theta_i^{(k)}, \theta_i^{(k+1)}) = \min\left\{\frac{h(\theta_i^{(k+1)})}{h(\theta_i^{(k)})}, 1\right\}. \quad (14)$$

Suppose the Markov chain is currently at $\{\theta^{(k)}, \beta^{(k)}, \sigma^{2(k)}\}$. To move to the next step $\{\theta^{(k+1)}, \beta^{(k+1)}, \sigma^{2(k+1)}\}$, we proceed as follows: (i) For $i=1, \dots, m$, to update θ_i , we draw a candidate $\theta_i^{(k+1)}$ from the log-normal density $f(\theta_i | \beta^{(k)}, \sigma^{2(k)})$. With probability $\alpha(\theta_i^{(k)}, \theta_i^{(k+1)})$ given by (14), this candidate is accepted, otherwise set $\theta_i^{(k+1)} = \theta_i^{(k)}$; (ii) To update β , draw $\beta^{(k+1)}$ from $[\beta | Y, \theta^{(k+1)}, \sigma^{2(k)}]$; (iii) To update σ^2 , draw $\sigma^{2(k+1)}$ from $[\sigma^2 | Y, \theta^{(k+1)}, \beta^{(k+1)}]$.

The estimation of $E(\theta_i | Y)$ and $V(\theta_i | Y)$ is based on the marginal sample $\{\theta_i^{(k)}\}$ from the Gibbs sampler. We will compare this direct Gibbs sampling approach with the logarithmic transformation approach through a numerical study in Section 5.

5. A NUMERICAL STUDY

We now compare the two approaches introduced in Sections 3 and 4, i.e., the transformation approach and the complete HB approach, for the area-level nonlinear model given by (6) and (7) through a small simulation study. We are mainly interested in the relative differences of the posterior estimates based on the two approaches. For this, we consider the following nonlinear mixed random effect model with $m=12$ small areas:

$$y_i \sim N(\theta_i, \theta_i \psi_i^2), \quad i = 1, \dots, 12; \quad (15)$$

and

$$\log(\theta_i) = \beta_0 + x_{1i} \beta_1 + x_{2i} \beta_2 + v_i, \quad v_i \sim N(0, \sigma^2); \quad (16)$$

where the x_{1i} and x_{2i} are set as: $x_1 = (1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0)'$ and $x_2 = (-6, -5, -4, -3, -2, -1, 1, 2, 3, 4, 5, 6)'$. The fixed effects coefficients are set at $\beta_0 = -1.5, \beta_1 = 1.0, \beta_2 = 0.5$. Variance component σ^2 was set at 1.5, and ψ_i^2 was chosen to be $\psi_i^2 = 1/n$, where n in practice is usually considered to be the sample size for the i th small area. In our simulation, n was chosen to be 5, 10, 20, 50, and 100.

For each n , we generated 10 data sets according to models (15) and (16). Let $\hat{\theta}_i^T$ denote the estimator of the posterior mean of θ_i based on the transformation approach, and $\hat{\theta}_i^B$ denote the corresponding estimator based on the complete HB approach. For both

approaches, the Gibbs sampler was run for 10,000 iterations with the first 5,000 iterations as “burn-in” period, and the last 5,000 iterations were kept for analysis. For each data set, we calculated $\hat{\theta}_i^T$ and $\hat{\theta}_i^B$, and the corresponding posterior variances $\hat{V}^T(\theta_i)$ and $\hat{V}^B(\theta_i)$. To compare these estimators, we calculated the relative difference (RD) of the two estimators $\hat{\theta}_i^T$ and $\hat{\theta}_i^B$ as $RD_i = |\hat{\theta}_i^B - \hat{\theta}_i^T| / \hat{\theta}_i^B$. We also calculated the ratio of the two posterior variances as $RV_i = \hat{V}^B(\theta_i) / \hat{V}^T(\theta_i)$.

Table 1 presents the summarized measures of RD and RV based on the 10 data sets over the 12 synthetic small areas. The summary measures include the minimum (min), maximum (max) and mean over the 12 small areas.

For each area, as n increases, or as ψ_i^2 decreases, the relative difference becomes smaller. In particular, when $n=100$, or $\psi_i^2 = 0.01$, the RD_i is between 0.45% to 0.99%, the average over the areas is only 0.68%, even less than 1%. When $n=20$ or $\psi_i^2 = 0.05$, the RD_i is between 2.81% to 4.42%, the average is 3.16%, which is less than 5%. On the other hand, when $n=5$ or $\psi_i^2 = 0.2$, the relative difference is quite large, ranging from 9.59% to 27.68%, and the average is about 18.4%.

For the posterior variance, Table 1 shows that when $n=100$, or $\psi_i^2 = 0.01$, the two posterior variances, $\hat{V}^T(\theta_i)$ and $\hat{V}^B(\theta_i)$, are very close to each other, with the ratio $\hat{V}^B(\theta_i) / \hat{V}^T(\theta_i)$ ranging from 0.993 to 1.135, and the average is about 1.045. When $n=20$, the ratio ranges from 1.413 to 1.894, and the average is 1.593. When $n=5$, the ratio could be very large, and the average is about 6.270.

In summary, if $\psi_i^2 = 0.01$ or the CV of y_i is 10%, then the two approaches are approximately equivalent to each other with approximately equal posterior mean estimators and posterior variance estimators, i.e., $\hat{\theta}_i^T \approx \hat{\theta}_i^B$ and $\hat{V}^T(\theta_i) \approx \hat{V}^B(\theta_i)$. If $\psi_i^2 = 0.02$ or the CV of y_i is about 14.14%, $\hat{V}^T(\theta_i)$ is about 87.27% of $\hat{V}^B(\theta_i)$. If the CV of y_i is about 22.36%, then $\hat{V}^T(\theta_i)$ is about 62.77% of $\hat{V}^B(\theta_i)$. Thus in practice, care should be taken when using logarithmic transformation on the original data.

When the CV of y_i is less than 10%, the logarithmic transformation approach works quite well. If the CV is about 15%, the logarithmic transformation could lead to about 15% underestimation of the true posterior variance obtained from the complete Gibbs sampling approach, even though the posterior mean estimates are about the same. However, if the CV of y_i is large, say over 20%, the transformation approach could lead to serious underestimation of the posterior variance and relatively large bias in the posterior mean estimator.

6. CONCLUSION REMARKS

In this paper, we have studied a nonlinear mixed effects area level model for small area estimation using a log-transformation approach and a complete HB approach. Our results have shown that log-transformation approach could lead to estimation bias and underestimation of variance. In general, a complete HB approach should be employed for small area estimation when using nonlinear mixed effects models. An interesting application of the nonlinear mixed effects model to the Canadian census undercoverage estimation can be found, for example, in You (2000).

Table 1. Comparison of RD_i and RV_i

n	CV(y_i)	RD_i %			RV_i		
		min	max	mean	min	max	mean
n=5	0.45	9.59	27.68	18.39	2.13	9.95	6.27
n=10	0.32	5.32	14.25	8.84	1.73	6.55	3.12
n=20	0.22	2.81	4.42	3.16	1.41	1.89	1.59
n=50	0.14	0.91	1.65	1.12	1.10	1.19	1.15
n=100	0.10	0.45	0.99	0.68	0.99	1.13	1.05

ACKNOWLEDGEMENT

This work was partially supported by a research grant to J.N.K. Rao from the NSERC of Canada.

REFERENCES

- Chip, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings Algorithm. *The American Statistician*, 49, 327-335.
- 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.
- Gelfand, A.E. and Smith, A.F.M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398-409.
- Ghosh, M. and Rao, J.N.K. (1994). Small area estimation: an appraisal (with discussion). *Statistical Science*, 9, 55-93.
- Hobert, J.P. and Cassella, G. (1996). The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association*, 91, 1461-1473.
- Rao, J.N.K. (1999). Some recent advances in model-based small area estimation. *Survey Methodology*, 25, 175-186.
- You, Y. (1999). *Hierarchical Bayes and Related Methods for Model-Based Small Area Estimation*. Ph.D. Thesis, Carleton University, Ottawa, Canada.
- You, Y. (2000). A nonlinear hierarchical modelling approach for census undercoverage estimation. *Proceedings of the Survey Methods Section, Statistical Society of Canada*, Ottawa.