Hierarchical Bayes Modeling of Survey-Weighted Small Area Proportions

Benmei Liu[1], Partha Lahiri[2] and Graham Kalton[1]
Benmei Liu, Westat, 1650 Research Blvd., Rockville, Maryland 20850[1]
University of Maryland, College Park, Maryland 20742[2]

Abstract

When a Hierarchical Bayes area level model is used to produce estimates of proportions of units with a given characteristic for small areas, it is commonly assumed that the survey weighted proportion for each sampled small area has a normal distribution and that the sampling variance of this proportion is known. However, these assumptions are problematic when the small area sample size is small or when the true proportion is near 0 or 1. In an effort to overcome these problems, we test two alternative models for the survey weighted proportion using a Monte Carlo simulation study in which stratified simple random samples are generated from a fixed finite population. We compare the results obtained from these alternative models with those obtained from two commonly used models.

Keywords: Weighted proportions, Hierarchical Bayes modeling, beta distribution

1. Introduction

Small area estimation methods are often used to estimate the proportions of units with a given characteristic for small areas. For example, small area estimation methods are used: in the Census Bureau's Small Area Income and Poverty Estimates (SAIPE) program to estimate poverty rates for states, counties, and school districts (Citro and Kalton, 2000; Maples and Bell, 2005); with data from the National Survey on Drug Use and Health (NSDUH) to estimate substance rates for states (Wright et al., 2007); and with data from the National Assessment of Adult Literacy (NAAL) to estimate proportions at the lowest level of literacy for states and counties (Mohadjer et al., 2007). In each case, the survey's sample sizes in the small areas are not large enough to support direct estimates of adequate precision. A wide variety of methods has been developed to address such small area estimation problems. See Rao (2003) and Jiang and Lahiri (2006a) for reviews, and Chattopadhyay et al. (1999), Farrell et al. (1997), and Malec et al. (1997, 1999) for methods specifically for estimating small area proportions. The range of methods includes both empirical best prediction (EBP) and Hierarchical Bayes (HB) approaches and models developed at both the area and unit levels. The approach used here is restricted to HB area level models.

When a Hierarchical Bayes area level model is used to produce estimates of proportions of units with a given characteristic for small areas, it is commonly assumed that the survey-weighted proportion for each sampled small area has a normal distribution and that the sampling variance of this proportion is known. However, these assumptions are problematic when the small area sample size is small or when the true proportion is near 0 or 1. In an effort to overcome these problems, we propose two alternative models for small area proportions and compare them with two commonly used models. The models are described in Section 3. The four models are compared by means of a Monte Carlo simulation study in which stratified simple random samples are generated from a fixed finite population. The simulation study is described in Section 4 and the results are presented in Section 5. The paper finishes with some concluding remarks in Section 6. First, however, we introduce the notation for a stratified simple random sample design in Section 2.

2. Notation

Let N_{ih} denote the population size in stratum h in area i of a finite population ($i = 1,...,m; h = 1,...,H_i$). Let y_{ihk} be the binary response for the characteristic of interest for unit k in stratum h in area i ($k = 1,...,N_{ih}$). The parameters to be estimated are the small area proportions $P_i = \Sigma_h \Sigma_k y_{ihk} / N_{hk}$.

With the stratified simple random sample design under study, n_{ih} units are selected from the N_{ih} units in stratum (ih). The standard direct survey estimator is:

$$p_{iw} = \frac{\Sigma_h^{H_i} \Sigma_k^{n_{hi}} w_{ih} y_{ihk}}{\Sigma_h^{H_i} \Sigma_k^{n_{hi}} w_{ih}}, i = 1,...,m \qquad (2.1)$$

where w_{ih} denotes the sampling weight given by $w_{ih} = N_{ih}/n_{ih}$.

The variance of p_{iw} can be expressed as

$$VAR_{st}(p_{iw}) = \frac{P_i(1-P_i)}{n_i} DEFF_i \qquad (2.2)$$

where $DEFF_i$ is the design effect reflecting the effect of the complex sample design (Kish, 1965). For a stratified simple random sample with a negligible sampling fraction in all strata, the design effect is given approximately by:

$$DEFF_i = \frac{\Sigma_h W_{ih}^2 P_{ih}(1-P_{ih})/n_{ih}}{P_i(1-P_i)/n_i} \qquad (2.3)$$

where $W_{ih} = N_{ih}/N_i$, $N_i = \Sigma_h N_{ih}$, $n_i = \Sigma_h n_{ih}$, and P_{ih} is the population proportion in stratum h in area i.

The design effect $DEFF_i$ is a function of the P_{ih}, which are unknown. If $P_{ih} \approx P_i$, $DEFF_i$ can be approximated by $deff_{iw} = n_i \Sigma_h W_{ih}^2 / n_{ih}$. This approximation is known and can be easily computed from the data.

The problem with p_{iw} is that it is very imprecise when the sample size n_i is small. Small area estimation procedures can be used to address this problem. Section 3 describes the HB area level models investigated in this study.

3. Models Studied

A general area level model consists of two models. One—the sampling model—is a model for the sampling error of the direct survey estimates. The other—the linking model—relates the population value for an area-to-area specific auxiliary variables $x_i = (x_{i1},...,x_{ip})'$.

Section 3.1 describes two area models that are often used for estimating small area proportions and Section 3.2 outlines some problems associated with these models. Section 3.3 describes two alternative models that may serve to address these problems.

3.1 Two Commonly Used Models

We study two commonly used models for comparison with the new models described in Section 3.4. The first is the Fay-Herriot model (Fay and Herriot, 1979), which assumes known sampling variances and normal distributions for both the sampling and the linking models. The second is the normal-logistic model, which differs from the Fay-Herriot model only by the replacement of a logit-normal distribution for the normal distribution in the linking model.

Model 1: (Fay-Herriot normal-normal model)

Sampling model:

$$p_{iw} \mid P_i \sim N(P_i,\ \psi_i) \qquad (3.1)$$

Linking model:

$$P_i \mid \beta, \sigma_v^2 \sim N(x_i'\beta, \sigma_v^2) \qquad (3.2)$$

Model 2: (normal-logistic model)

Sampling model:

$$p_{iw} \mid P_i \sim N(P_i,\ \psi_i) \qquad (3.3)$$

Linking model:

$$g(P_i) \mid \beta, \sigma_v^2 \sim N(x_i'\beta, \sigma_v^2) \qquad (3.4)$$

In both models the sampling variance ψ_i is assumed to be known. Model 1 is referred as a matched model because the sampling and linking models can be combined to produce a relatively simple linear mixed model. However, a nonlinear linking model is often preferred for modeling proportions, leading to unmatched sampling and linking models, as in Model 2 (see, for example, You and Rao, 2002). The link function $g(\bullet)$ can be empirically determined by checking the model fit. The *log* and *logit* link functions have been used. The $logit(P_i)$ linking model is chosen here in order to guarantee that the estimate of P_i always fall into the right range of (0, 1).

3.2 Issues with Model 1 and 2

There are two main issues associated with Models 1 and 2. The first is that both models assume

known sampling variances ψ_i, whereas in practice they have to be estimated. A simple approach is to use the direct variance estimate but that estimate is very imprecise when P_i is either very small or very large and when the sample size n_i is small. An alternative, more complex, approach is to develop an approximate estimate of P_i, say p_{isyn}, from a simple model such as a logistic model for p_{iw} in terms of the auxiliary variables, and then use that estimate in the following synthetic variance estimator:

$$\text{var}_{stsyn} = \frac{p_{isyn}(1 - p_{isyn})}{n_i} deff_{iw} \qquad (3.5)$$

When there are no auxiliary variables available, the overall sample proportion may be used for p_{isyn} in the computation of the synthetic variance estimator.

The second issue concerns the normality assumption in the sampling model, which is based on a large sample approximation. When the sample size n_i is small and P_i is near 0 or 1, as is often the case with small area estimation, that approximation is problematic.

3.3 Two Alternative Models

Under Models 1 and 2, the unknown sampling variances ψ_i are estimated in some way, and then the resultant estimates are treated as if they were the known true values. A possible alternative approach is to treat the ψ_i as unknown parameters in the HB model. This approach is adopted in Model 3, as a variant of Model 2.

A possible approach for addressing the nonnormality of the sampling distributions of the survey-weighted small area proportions is to replace the normal distribution assumption by an alternative distribution. That approach is applied in Model 4 with the assumption of a beta sampling distribution; a distribution that has the desirable property of having a (0, 1) range. In other regards Model 4 is the same as Model 3, including treating the ψ_i as unknown parameters. Model 4 was initially considered by Jiang and Lahiri (2006b) for an EBP approach in one of their illustrative examples to estimate finite population domain means.

Model 3 (normal-logistic model with unknown sampling variance):

Sampling model:

$$p_{iw} \mid P_i \sim N(P_i,\ \psi_i) \qquad (3.6)$$

Linking model:

$$logit(P_i) \mid \beta, \sigma_v^2 \sim N(x_i'\beta, \sigma_v^2) \qquad (3.7)$$

Model 4: (beta-logistic model with unknown sampling variance)

Sampling model:

$$p_{iw} \mid P_i \sim beta(a_i,\ b_i) \qquad (3.8)$$

Linking model:

$$logit(P_i) \mid \beta, \sigma_v^2 \sim N(x_i'\beta, \sigma_v^2) \qquad (3.9)$$

For both Model 3 and Model 4, the approximate variance function $\psi_i = [P_i(1 - P_i)/n_i]deff_{iw}$ is used. The parameters a_i and b_i in Model 4 are then given by:

$$a_i = P_i\left(\frac{n_i}{deff_{iw}} - 1\right), \text{ and } b_i = (1 - P_i)\left(\frac{n_i}{deff_{iw}} - 1\right).$$

HB small area estimates can be computed from all four models using the Metropolis-Hasting algorithm within the Gibbs sampler. Details of the algorithm, which draws random samples based on the full conditional distributions of the unknown parameters starting with one or multiple sets of initial values, are given by Robert and Casella (1999) and Chen, Shao, and Ibraham (2000).

4. Simulation Study

4.1 The Study Population and the Sample Design

This section describes the simulation study that was conducted to compare the efficiency of the small area estimates produced by the four HB models. The simulation study was based on the 2002 Natality public-use data file. The file included all births occurring within the United States in 2002. Data were obtained from certificates filed for births occurring in each State. Details about the births recorded in the National Vital Statistics System are given at the

website for the National Center for Health Statistics (http://www.cdc.gov/nchs/births.htm).

The finite population studied comprised 4,024,378 records of live births in the U.S. with birth weights reported. The parameter of interest is the state level low birth weight rate P_i, $i = 1,...,51$, where low birth weight is defined as less than 2,500 grams. The value of P_i varied from 5 percent to 11 percent across the states.

Within each state, a stratified SRS design was used to draw samples from the birth records. Mother's race (White, Black, and Others) was used as the stratification variable. The national sample size was set to be about 1,500 birth records for each race group. A uniform sampling fraction was used across the states for each race group, subjecting to the condition at least two birth records were sampled within each race group in each state. The resultant national sample size turned out to be $n = 4{,}526$ birth records. The state level sample sizes n_i ranged from 7 (for small states such as Vermont) to 690 (for California), with a median sample size of 61. This sampling procedure was repeated $R = 1{,}000$ times, creating 1,000 independent sample data sets. The sampling weights remained the same over different simulation runs.

4.2 Computation of the HB Estimates

For simplicity, the following assumptions were made for the HB models:

1. No auxiliary variables were used, so that $x_i'\beta = \mu$;

2. For Models 1 and 2, the sampling variances were taken to be given by $\psi_i = (p_w(1 - p_w)/n_i)deff_{iw}$, where $p_w = \Sigma\Sigma\Sigma w_{ih} y_{ihk} / \Sigma\Sigma n_i w_{ih}$ is the national estimate of the proportion of low birth weigh live births. (A check on the use of the approximate $deff_{iw}$ in place of $DEFF_i$ showed that the approximation was reasonable: the two quantities were close, with a product moment correlation of 0.96 and a ratio of 1.08 for the means of the two).

3. Flat prior for μ, i.e., $f(\mu) \propto 1$, and inverse gamma for σ_v^2, i.e., $\sigma_v^2 \sim IG(0.001,\ 0.001)$.

For each sample data set, the first step in the computations was to calculate the state direct sample estimates. The estimates for each sample data set were then used in turn as input to the WinBUGS software, which was used to produce the HB estimates for all four models.

In a sizable number of the states with small n_i, the direct estimates were zero in some sample data sets. Since WINBUGS can handle direct estimates of zero only for Model 1, the zero direct estimates were perturbed to very small positive numbers for the other models.

For each WinBUGS run, three independent chains were used. For each chain, burn-ins of 10,000 samples were produced, with 10,000 samples after burn-in. The samples after burn-in were thinned by a factor of two to reduce auto-correlation of the MCMC. The resultant 15,000 MCMC samples after burn-in were then used to compute the posterior mean and percentiles for each HB model based on each sample data set. The potential scale reduction factor $\hat{R}$ was used as the primary measure for convergence (see Gelman and Rubin, 1992).

5. Simulation Results

Let P_i^{HB} denote an HB estimator of P_i, the percentage of low birth weight live births in state *i*, and let $P_{i,q}^{HB}$ denote the q^{th} percentile of the posterior distribution of P_i. Based on results from the 1,000 simulation data sets, Tables 1 and 2 on the next page present for each model: the noncoverage probability for the 95 percent credible intervals, i.e., the probability that the interval from $P_{i,.025}^{HB}$ to $P_{i,.975}^{HB}$ fails to cover P_i; the mean width of the credible intervals $P_{i,.975}^{HB} - P_{i,.025}^{HB}$; and standard deviations of the credible interval widths.

To examine the effect of state sample size on the simulation results, the 51 states are placed in 3 groups according to their sample size: small $(n_i \le 30)$; medium $(30 < n_i \le 100)$; and large $(n_i > 100)$. The results presented in the tables are overall averages across all states and averages for the three groups separately.

Table 1 reports the average percentage of times that the 95 percent credible interval for each P_i failed to cover the true value of P_i over the 1,000

replications. The upper half of Table 2 displays the average widths of the 95 percent credible intervals and the lower half of the table presents the standard deviations of these widths. The Fay-Herriot model (M1) credible intervals are very conservative, giving nearly zero noncoverage. This result is obtained at the cost of the largest average credible interval width among the four models. The M1 credible interval widths are very stable. A small proportion of the M1 credible intervals had negative lower bounds.

A possible explanation for the low level of noncoverage with M1 is that the sampling variances were overestimated, perhaps because $deff_{iw}$ was used for $DEFF_i$. To examine this possibility, we used $DEFF_i$ in computing the sampling variance and found virtually no difference in the noncoverage rate. We also ran the model with the true variance as defined in (1.2) and again found no appreciable difference in the noncoverage rate. The nonnormality of the sampling distribution of p_{iw} could also be a source of this problem.

At 8.2 percent, the noncoverage rate of the credible intervals for the normal-logistic model (M2) is above the nominal rate of 5 percent. This model has the smallest average interval width. The noncoverage rate for the normal-logistic model with unknown variance (M3) is closer to the nominal rate, with an average interval width that is somewhat larger than that for M2.

The noncoverage rate for the beta-logistic model (M4) of 4.4 percent overall is closest to the nominal noncoverage rate. However, the average width of the credible intervals is larger than those for M2 and M3 and the variability of the interval width is larger than that of the other three models. This instability may be due to the complexity of the full conditional distribution for the beta model. The large proportion of the 1,000 direct estimates that were 0 for some of the small states may also cause significant problems in fitting the beta distribution.

As is to be expected, for all four models the mean width of the credible intervals declines with increasing state sample size and the variation in the widths also declines with increased sample size. Despite these declines, however, the noncoverage rates also decline with increasing sample size for Models 2, 3, and 4. The noncoverage rates are in fact very small for the states with large n_i, suggesting that the credible intervals are not adequately reflecting the effect of the greater precision of the direct estimates in the states with large sample sizes.

Table 1. Percentage of times that the 95 percent credible intervals fail to cover the state parameters based on the 1,000 simulations

Sample size	M1	M2	M3	M4
Overall	0.4	8.2	6.5	4.4
Small n_i	0.1	10.8	8.0	6.3
Medium n_i	0.5	9.8	7.9	4.4
Large n_i	0.7	1.9	1.9	1.7

Table 2. Mean 95 percent credible interval width and mean standard deviation of the 95 percent credible interval widths based on 1,000 simulations (in percentages)

Sample size	M1*	M2	M3	M4
Mean width				
Overall	9.0	5.5	6.2	8.5
Small n_i	10.2	5.9	6.8	9.2
Medium n_i	9.1	5.6	6.3	8.7
Large n_i	7.3	4.8	5.3	6.9
Mean standard deviation				
Overall	0.8	1.9	2.1	3.1
Small n_i	1.1	2.4	2.6	4.1
Medium n_i	0.8	1.9	2.1	3.1
Large n_i	0.5	1.2	1.4	1.8

* Note: For Model 1, a small proportion of the credible intervals had negative lower bounds.

6. Discussion

In the simulation study, we have compared design-based coverage properties of credible intervals resulting from different hierarchical Bayes models for estimating small area proportions from a stratified simple random sample design. The hierarchical Bayes version of the well-known Fay-Herriot model appears to produce overly conservative credible intervals. The non-normality of both the sampling and the linking models is a possible source of this problem.

The credible intervals for the beta-logistic hierarchical model achieve almost the nominal coverage for the finite population proportions.

However, since one of the full conditionals for the beta-logistic model involves the survey-weighted proportions, there is a problem with the MCMC whenever the survey-weighted proportion is zero. The credible intervals for this model are also wider than those for the other two models with a logistic linking model. It may be possible to reduce the width of the credible interval for the beta-logistic model by modifying the model in some way, such as by employing a suitable two-part random effects model that will avoid the problem of survey-weighted proportions of zero. We plan to undertake this research in the future.

The simulation study was restricted to a very simple sample design. In addition, for simplicity no auxiliary variables were included in the linking models, whereas in practice the inclusion of such variables is routine and almost essential. We plan to conduct further simulation studies to cover different sample designs, different sample sizes, and to incorporate some auxiliary variables in the linking models.

References

Chattopadhyay, M., Lahiri, P., Larsen, M., and Reimnitz, J. (1999). Composite estimation of drug prevalence for sub-state areas. *Survey Methodology*, 25, 81-86.

Chen, M., Shao, Qi., and Ibraham, J.G. (2000). *Monte Carlo methods in Bayesian computation.* New York: Springer-Verlag.

Citro, C., and Kalton, G. (Eds.). (2000). *Small-Area Income and Poverty Estimates: Priorities for 2000 and Beyond*. Washington, DC: National Academy Press.

Farrell, P.J., MacGibbon, B., and Tomberlin, T.J. (1997). Empirical Bayes estimators of small area proportions in multistage designs. *Statistical Sinica*, 7, 1065-1083.

Fay, R.E., and Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedure to census data. *Journal of the American Statistical Association*, 74, 269-277.

Gelman, A., and Rubin, D.B. (1992). Inference from iterative simulation using multiple sequence. *Statistical Science*, 7, 457-472.

Jiang, J., and Lahiri, P. (2006a). Mixed model prediction and small area estimation. *Test*, 15, 111-999.

Jiang, J., and Lahiri, P. (2006b). Estimation of finite population domain means: A model-assisted empirical best prediction approach. *Journal of the American Statistical Association*, 101, 301-311.

Kish, L. (1965). *Survey sampling*. New York: John Wiley.

Malec, D., Davis, W., and Cao, X. (1999). Small area estimates of overweight prevalence using sample selection adjustment. *Statistics in Medicine*, 18, 3189-3200.

Malec, D., Sedransk, J., Moriarity, C.L., and Lecler, F.B. (1997). Small area inference for binary variables in National Health Interview Survey. *Journal of the American Statistical Association,* 92, 815-826.

Maples, J., and Bell, W.R. (2005). Evaluation of school district poverty estimates: Predictive models using IRS income tax data*. Proceedings of the American Statistical Association.*

Mohadjer, L., Rao, J.N.K., Liu, B., Krenzke, T., and VanDekerckhove, W. (2007). Hierarchical Bayes small area estimates of adult literacy using unmatched sampling and linking models. *Proceedings of the American Statistical Association.*

Rao, J.N.K. (2003). *Small area estimation*. New York: John Wiley and Sons.

Robert, C.P., and Casella, G. (1999). *Monte Carlo statistical methods*. New York: Springer-Verlag.

Wright, D., Sathe, N., and Spagnola, K. (2007). *State estimates of substance use from the 2004-2005 National Surveys on Drug Use and Health.* (DHHS Publication No. SMA 07-4235, NSDUH Series H-31). Rockville, MD: Substance Abuse and Mental Health Services Administration, Office of Applied Studies.

You, Y., and Rao, J.N.K. (2002). Small area estimation using unmatched sampling and linking models. *Canadian Journal of Statistics*, 30, 3-15.