Available online at www.isas.org.in/jisas
JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 66(1) 2012 121-124

Inferences on Small Area Proportions

Shijie Chen[1] and P. Lahiri[2*]
[1]*Bristol-Myers Squibb, 311 Pennington-Rocky Hill Road, Pennington, NJ 08534, USA*
[2]*Joint Program of Survey Methodology, University of Maryland, College Park, MD 20742, USA*
Received 02 May 2011; Revised 19 July 2011; Accepted 17 August 2011

SUMMARY

Design-based methods are generally inefficient for making inferences about small area proportions for rare events. In this paper, we discuss an alternative hierarchical model and the associated hierarchical Bayes methodology. Sufficient conditions for propriety of the posterior distributions of relevant parameters are presented.

Keywords : Credible interval, MCMC, Rare event.

1. INTRODUCTION

Estimation of finite population proportions for rare events in presence of small sample sizes received considerable interest over the years. Design-based methods given in standard text books (e.g., Cochran 1977) are not suitable for such estimation for number of reasons. For given sample, the proportion estimate as well as the associated standard error estimate often turns out to be zero since small sample fails to detect the rare event, giving misleading picture of the real situation. The true variability is expected to be high due to sample size and the associated normality-based confidence interval is likely to be wide and may also suffer from coverage error due to discreteness and asymmetry of the underlying probability distribution.

To address this problem, commonly referred to as the small area estimation problem, hierarchical models that combine information from different sources have been used in the past. A convenient way to implement such models is to use hierarchical Bayes methodology using Monte Carlo Markov Chain (MCMC). Using design-based Monte Carlo simulation, Liu *et al*. (2007) evaluated design-based properties of hierarchical Bayes credible intervals for a number of area level models where the sampling distribution of survey-weighted proportions is assumed to be continuous distribution as in the case of the well-known Fay-Herriot model (see Fay and Herriot 1979). They noted the difficulty in modeling the design-based estimates of proportions by a continuous distribution since the sampling distribution is inherently discrete due to small sample and binary nature of the observations. An alternative approach would be to model the binary observations. Such methods are described in Rao (2003) and Jiang and Lahiri (2006). In this paper, we provide hierarchical Bayes methodology that can be used to produce small area estimates of proportions for rare events and present sufficient conditions for the propriety of the relevant posterior distributions.

2. HIERARCHICAL BAYES METHOD

Let y_{ij} (0 or 1) be the observation for the jth individual in the ith small area ($i = 1, ..., m$; $j = 1, ..., N_i$). Our interest is to make inference about the finite population proportion $\bar{Y}_i = \sum_{j=1}^{N_i} y_{ij} / N_i$, where N_i is

[*] *Corresponding author* : P. Lahiri
E-mail address : plahiri@survey.umd.edu

the known population size of the *i*th small area (i = 1, ..., m). We assume that the following model adequately describes the finite population.

Model

(i) $y_{ij} \mid \theta_i \overset{ind}{\sim}$ Bernoulli $\left(\frac{\exp(\theta_i)}{1+\exp(\theta_i)}\right)$

(ii) $\theta_i \overset{ind}{\sim} \mathrm{N}(x_i'\delta, r^{-1})$

(iii) $\delta \sim$ Uniform(R^p) independent of $r \sim$ Gamma (a/2, b/2)

where a and b are known and $x_i' = (x_{i1}, ..., x_{ip}) \in R^p$ (i = 1, ..., m; j = 1, ..., N_i). In different applications, small values of a and b are often recommended because such choice produces vague prior on r. See Malec *et al*. (1997) for a similar model for estimation of small-area proportions of doctor's visits using National Center for Health Statistics data.

We assume a non-informative sample design. Let s be the set of sampled units and $y_s = \{y_{11}, ..., y_{1n_1}, ..., y_{m1}, ..., y_{mn_m}\}'$. Inferences on $\bar{Y}_i$ will be based on the posterior distribution of $\bar{Y}_i$, which can be computed using the Monte Carlo Markov Chain (MCMC) method. To calculate the Bayes estimator (the mean of the posterior distribution of $\bar{Y}_i$) it is equivalent to calculate for i = 1, ..., m,

$$\mathrm{E}\left(\sum_{j=1}^{N_i} y_{ij} \mid y_s\right) = \sum_{j\in s} y_{ij} + \sum_{j\notin s} \mathrm{E}(y_{ij} \mid y_s).$$

Now for $j \notin$ s,

$$\mathrm{E}(y_{ij} \mid y_s) = \mathrm{E}(\mathrm{E}(y_{ij} \mid \theta_i, y_s) \mid y_s) = \mathrm{E}\left(\frac{\exp(\theta_i)}{1+\exp(\theta_i)} \mid y_s\right),$$

the posterior mean of $\pi_i = \exp(\theta_i)/[1 + \exp(\theta_i)]$. The posterior mean and variance of $\bar{Y}_i$ are given by

$$\mathrm{E}\left(\bar{Y}_i \mid y_s\right) = \frac{1}{N_i}\left(\sum_{j\in s} y_{ij} + (N_i - n_i)\,\mathrm{E}(\pi_i \mid y_s)\right) \qquad (1)$$

and

$$\mathrm{V}\left(\bar{Y}_i \mid y_s\right) = \mathrm{V}\left(\mathrm{E}\left(\bar{Y}_i \mid y_s, \theta_i\right) \mid y_s\right) + \mathrm{E}\left(\mathrm{V}\left(\bar{Y}_i \mid y_s, \theta_i\right) \mid y_s\right)$$

$$= \left(\frac{N_i - n_i}{N_i}\right)^2 \mathrm{V}(\pi_i \mid y_s) + \left(\frac{N_i - n_i}{N_i^2}\right) \mathrm{E}(\pi_i(1 - \pi_i) \mid y_s). \qquad (2)$$

Since the prior on δ is improper, we need conditions under which the posterior distribution of θ_i is proper. The following theorem ensures propriety under certain conditions.

Theorem : Denote $D = \{i \mid 1 \le y_{i.} \le n_i - 1\} := \{d_1, ..., d_q\} \subset \{1, ..., m\}$ where $d_1 < d_2 < ... d_q$, $X_D = (x_{d_1}, x_{d_2}, ..., x_{d_q})'$ and $y_{i.} = \sum_{j=1}^{n_{ij}} y_{ij}$. Assume that $b > 0$ and $m - p + a > 0$. If $X_D'X_D$ is nonsingular, the joint posterior probability distribution function of the θ_i's given y_s is proper.

Proof of Theorem: The joint probability distribution functions of $\theta = (\theta_1, ..., \theta_m)'$, δ and r given y_s for the following two cases are different and so we need to handle these two cases separately:

Case 1. $1 \le y_{i.} \le n_i - 1\ \forall y_{i.}$ (i = 1, ..., m).

Case 2. Assume that $q < m$ of the $y_{i.}$'s are neither 0 nor n_i (i = 1, ..., m).

Case 1

The joint probability distribution function of $\theta = (\theta_1, ..., \theta_m)'$, δ and r given y_s is given by

$$\pi(\theta, \delta, r \mid y_s) \propto \prod_i \prod_{j=1}^{n_i}\left(\frac{\exp(\theta_i)}{1+\exp(\theta_i)}\right)^{y_{ij}} \left(\frac{1}{1+\exp(\theta_i)}\right)^{1-y_{ij}} \times \prod_i r^{\frac{1}{2}} \exp\left(-\frac{r}{2}(\theta_i - x_i'\delta)^2\right) \times r^{\frac{a}{2}-1} \exp(-br/2).$$

Since

$$\sum_i (\theta_i - x_i'\delta)^2 = \delta'\left(\sum_i X_i X_i'\right)\delta - 2\delta'\sum_i X_i\theta_i + \sum_i \theta_i^2$$
$$= \delta'(X'X)\delta - 2\delta'(X'\theta) + \theta'\theta$$
$$= (\delta - (X'X)^{-1}X'\theta)'(X'X)(\delta - (X'X)^{-1}X'\theta) + \theta'(I - X(X'X)^{-1}X')\theta,$$

where $X = (x_1, x_2, ..., x_m)'$, integrating with respect to δ, we have

$$\pi(\theta, r \mid y_s) \propto \prod_i \frac{\exp(\theta_i y_{i.})}{(1+\exp(\theta_i))^{n_i}} \times r^{\frac{m+a-p}{2}-1} \times \exp\left[-\frac{r}{2}(b+\theta'(I-X(X'X)^{-1}X')\theta)\right].$$

Integrating with respect to r, we have

$$\pi(\theta \mid y_s) \propto \prod_i \frac{\exp(\theta_i y_{i.})}{(1+\exp(\theta_i))^{n_i}} \times \left[b+\theta'(I-X(X'X)^{-1}X')\theta\right]^{-\frac{m+a-p}{2}} \le c \times \prod_i \frac{\exp(\theta_i y_{i.})}{(1+\exp(\theta_i))^{n_i}}, \quad (3)$$

for some constant c since $Q := I - X(X'X)^{-1}X'$ is nonnegative definite. Thus the result follows from

$$\int_{-\infty}^{\infty} \frac{\exp(\theta_i y_{i.})}{(1+\exp(\theta_i))^{n_i}} d\theta_i = \int_0^1 p^{y_{i.}-1}(1-p)^{n_i-y_{i.}-1} dp < \infty.$$

Case 2

Without loss of generality, let $1 \le y_{1.} \le n_1 - 1$, ..., $1 \le y_{q.} \le n_q - 1$. Then $D = \{1, ..., q\}$ and $y_{i.} = 0$ or n_i for $q + 1 \le i \le m$ and

$$\int \dots \int \pi(\theta, \delta, r \mid y_s)\, dr d\delta \prod_{i=1}^m d\theta_i \propto \int \dots \int \prod_{i=1}^q \frac{\exp(\theta_i y_{i.})}{(1+\exp(\theta_i))^{n_i}} \times \prod_{i=1}^q r^{\frac{1}{2}} \exp\left(-\frac{r}{2}(\theta_i - x_i'\delta)^2\right) \prod_{i=1}^q d\theta_i \times \int \dots \int \prod_{i=q+1}^m \frac{\exp(\theta_i y_{i.})}{(1+\exp(\theta_i))^{n_i}} \times \prod_{i=q+1}^m r^{\frac{1}{2}} \exp\left(-\frac{r}{2}(\theta_i - x_i'\delta)^2\right) \prod_{i=q+1}^q d\theta_i \times r^{\frac{a}{2}-1} \exp(-br/2) dr d\delta.$$

Now,

$$\int \dots \int \prod_{i=q+1}^m \frac{\exp(\theta_i y_{i.})}{(1+\exp(\theta_i))^{n_i}} \times \prod_{i=q+1}^m r^{\frac{1}{2}} \exp\left(-\frac{r}{2}(\theta_i - x_i'\delta)^2\right) \prod_{i=q+1}^m d\theta_i \le \int \dots \int \prod_{i=q+1}^m r^{\frac{1}{2}} \exp\left(-\frac{r}{2}(\theta_i - x_i'\delta)^2\right) \prod_{i=q+1}^m d\theta_i = (2\pi)^{\frac{m-q}{2}}.$$

Hence,

$$\int \dots \int \pi(\theta, \delta, r \mid y_s) dr d\delta \prod_{i=1}^m d\theta_i \le (2\pi)^{\frac{m-q}{2}} \int \dots \int \prod_{i=1}^q \frac{\exp(\theta_i y_{i.})}{(1+\exp(\theta_i))^{n_i}} \times \prod_{i=1}^q r^{\frac{1}{2}} \exp\left(-\frac{r}{2}(\theta_i - x_i'\delta)^2\right) \prod_{i=1}^q d\theta_i \times r^{\frac{a}{2}-1} e^{-\frac{b}{2}r}\, dr d\delta < \infty$$ (follows from Case 1.)

Corollary: If $b = 0$ and $a > p + 1 - m$, the joint posterior probability distribution function of the θ_i's given y_s is improper.

Proof of Corollary

Let $\Omega := \left\{\theta : \sum_{i=1}^m \theta_i^2 \le K\right\}$, where K is a fixed positive number. From (3) and noting that there exits a positive constant c_0 such that

$$0 < c_0 < \frac{\exp(\theta_i y_{i.})}{(1+\exp(\theta_i))^{n_i}} < 1 \ \forall \ \theta \in \Omega,$$

we have in Ω,

$$c_1 \cdot [\theta' Q \theta]^{-\frac{m+a-p}{2}} \le \pi(\theta \mid y_s) \le c_2 \cdot [\theta' Q \theta]^{-\frac{m+a-p}{2}},$$

where c_1 and c_2 are positive constants. Hence, noting that Q is nonnegative definite with rank $m - p$, if $a > p + 1 - m$, we have

$$\int_\Omega [\theta' Q \theta]^{-\frac{m+a-p}{2}} d\theta = +\infty$$

and impropriety follows.

Remark 1. For the well-known prior on r in which $g(r) \propto r^\alpha$, $\forall \ \alpha > (p - 1 - m)/2$, or equivalently, for the prior on $A = \frac{1}{r}$, $h(A) \propto A^\beta$, $\forall \ \beta < (m + 3 - p)/2$, the posterior probability distribution function of the θ_i's given y_s is improper.

3. CONCLUDING REMARKS

The hierarchical model presented in the paper should work fairly well when we have exchangeability within the small areas. However, in many applications the assumption of exchangeability may not be reasonable because of the complexity in the inherent population structure. To deal with such complex situations, one may consider inclusion of unit level covariates. Also, the survey design may be informative, which poses additional complexity in the hierarchical Bayes methodology. The success of any hierarchical Bayes method, just like any other model-based method, depends on the underlying model. Thus the role of model selection and model diagnostics cannot be overemphasized. Finally, it will be instructive to evaluate design-based properties of the hierarchical Bayes procedures using Monte Carlo simulation similar to that given in Liu *et al*. (2007). These topics will be addressed in separate paper.

ACKNOWLEDGEMENTS

The second author's research was supported in part by NIH grant # R01 CA129101. The authors thank two anonymous referees for their constructive comments.

REFERENCES

Fay, R.E. and Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedure to census data. *J. Amer. Statist. Assoc.*, **74**, 269-277.

Jiang, J. and Lahiri, P. (2006). Mixed model prediction and small area estimation. *Test*, **15**, 1-96.

Liu, B., Lahiri, P. and Kalton, G. (2007). Hierarchical Bayes modeling of survey-weighted small area proportions. *Proceedings of the American Statistical Association*, Survey Research Section, 3181-3186.

Malec, D., Sedransk, J., Moriarity, C.L. and LeClere, F.B. (1997). Small area inference for binary variables in the national health interview survey. *J. Amer. Statist. Assoc.*, **92**, 815-826.

Rao, J.N.K. (2003). *Small Area Estimation*. Wiley Series in Survey Methodology. Wiley-Interscience, John Wiley Sons, Inc.