The Korean Journal of Applied Statistics (2009)
22(1), 153–161

Estimation of Small Area Proportions Based on Logistic Mixed Model

Kwang Mo Jeong[1] · Jung-Hyun Son[2]

[1]Dept. of Statistics, Pusan National University; [2]Consulting Division, ECMINER Co.

(Received November 2008; accepted December 2008)

Abstract

We consider a logistic model with random effects as the superpopulation for estimating the small area proportions. The best linear unbiased predictor under linear mixed model is popular in small area estimation. We use this type of estimator under logistic mixed model for the small area proportions, on which the estimation of mean squared error is also discussed. Two kinds of estimation methods, the parametric bootstrap and the linear approximation will be compared through a Monte Carlo study in the respects of the normality assumption on the random effects distribution and also the magnitude of sample sizes on the approximation.

Keywords: Best linear unbiased predictor, small area, logistic mixed model, mean squared error, parametric bootstrap.

1. Introduction

Sometimes we need to estimate the characteristics of small areas which are sub domains of the whole population. When the sample survey has been already performed for the whole population according to the sampling design we frequently encounter the presence of small sample sizes. Because the estimation of target parameters of small area based on its own sample is not satisfactory other statistical methods using auxiliary information from all the areas sharing common features are useful. Lohr and Prasad (2003) studied small area estimation using auxiliary information, and Ghosh and Rao (1994), Rao (2003) provide general discussions on small area estimation. The extra variations between small areas can further be explained by including the random effects across the small areas in a regression model. Even though we are to estimate additional variance components in the random effects regression model this approach is very popular in the small area estimation. Prasad and Rao (1990) theoretically discussed the best linear unbiased predictor(BLUP) and its mean squared error(MSE) under several types of linear regression model with random effects which is usually called a liner mixed model.

The proportions of certain attributes can be modeled effectively by a logistic regression model with random effects, which belongs to a family of generalized linear mixed models including the

This work was supported for two years by Pusan National University Research Grant.
[1]Corresponding author: Professor, Dept. of Statistics, Pusan National University, Busan 609-735, Korea. Email: kmjung@pusan.ac.kr

linear mixed model as a special case. Under the logistic mixed model we consider a BLUP type estimator for the small area proportions and we also discuss the estimation of MSE for the suggested estimator. González-Manteiga *et al.* (2007) applied Taylor series expansion to transform the logistic mixed model into an approximate linear mixed model and used the generalized least squares(GLS) estimator for the regression coefficients. On the other hand we may directly fit the logistic mixed model using the statistical softwares to obtain the maximum likelihood(ML) estimator for the regression coefficients and variance components. Numerical methods are used to approximate the likelihood function which includes integrals with respect to the density function of random effects distribution.

The mean squared error(MSE) is usually used as an accuracy measure of the BLUP as has been discussed by Kackar and Harville (1984), Prasad and Rao (1990). Under the assumption of normality for the random effects distribution the ML or the restricted maximum likelihood(REML) methods are used to estimate the variance components on which the MSE depends. Prasad and Rao (1990) provides estimators of MSE for the BLUP under several types of linear mixed models. But the estimation method cannot be directly applied to the BLUP type estimator of proportions under the logistic mixed model. In this situation the jackknife or the bootstrap approach is a good alternative to estimating the MSE. Lahiri (2003), Hall and Maiti (2006) discusses on the bootstrap estimation of MSE in various aspects of view.

In this paper we discuss the BLUP type estimator for the small area proportion and its MSE based on the logistic model with random intercepts. We compare the efficiency of MSE estimation methods between the linear approximation and the parametric bootstrap method by varying the distributions of random effects through a Monte Carlo study.

2. Logistic Model with Random Intercepts

2.1. Model for small area mean

Suppose a finite population of overall size N consists of D small areas of sizes N_d with $\sum_{d=1}^{D} N_d$. Let y_{dj} be the target variable of interest with a vector $\boldsymbol{x}_{dj}$ of k auxiliary variables for the j^{th} unit within d^{th} small area. Let $\boldsymbol{\delta} = \mathbf{A}\boldsymbol{y}$ be the parameter of interest for the population, where $\mathbf{A} = \text{diag}\{\boldsymbol{a}_d^t, d = 1, \ldots, D\}_{D \times N}$ and $\boldsymbol{a}_d^t = (a_{d1}, \ldots, a_{dN_d})$ with uniformly bounded known elements a_{dj}. When $a_{dj} = 1/N_d$ the parameter $\boldsymbol{\delta}$ corresponds to the mean vector $\bar{\boldsymbol{Y}}^t = (\bar{Y}_1, \ldots, \bar{Y}_D)$, where $\bar{Y}_d = \sum_{j=1}^{N_d} y_{dj}/N_d$ is the mean of d^{th} small area. Let S_d be a set of sampled units of size n_d from the d^{th} area with total sample size $n = \sum_{d=1}^{D} n_d$.

The linear mixed model for the target variable y_{dj} can be expressed as

$$y_{dj} = \boldsymbol{x}_{dj}^t \boldsymbol{\beta} + v_d + \epsilon_{dj}, \tag{2.1}$$

where $\boldsymbol{\beta}$ is a vector of fixed effects and v_d denotes the random effects of small areas and ϵ_{dj}'s are independent errors having variances σ_ϵ^2. When N_d is large the conditional mean of y_{dj} given v_d may be written as

$$\mu_d = \bar{\mathbf{X}}_d^t \boldsymbol{\beta} + v_d, \tag{2.2}$$

where $\bar{\mathbf{X}}_d$ is the vector of known means of the $\boldsymbol{x}_{dj}$ for the d^{th} area. According to Prasad and Rao (1990) the BLUP of μ_d is given by

$$\hat{\mu}_d(\boldsymbol{\theta}) = \bar{\mathbf{X}}_d^t \tilde{\boldsymbol{\beta}} + \gamma_d \left(\bar{y}_d - \bar{\boldsymbol{x}}_d^t \tilde{\boldsymbol{\beta}} \right), \tag{2.3}$$

where $\boldsymbol{\theta} = (\sigma_v^2,\ \sigma_\epsilon^2)^t$ and $\gamma_d = \sigma_v^2/(\sigma_v^2 + \sigma_\epsilon^2/n_d)$ for known values of σ_v^2 and σ_ϵ^2, and $\bar{\boldsymbol{x}}_d$ is the mean of $\boldsymbol{x}_{dj}$ for the sampled units in d^{th} area, and $\tilde{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^t\mathbf{V}^{-1}\boldsymbol{y}$ is a GLS estimator with the design matrix $\mathbf{X}$ and the covariance matrix $\mathbf{V}$ of $\boldsymbol{y}$.

We consider a finite population of small areas consisting of units generated from the superpopulation model (2.1). The BLUP for the small area mean $\bar{Y}_d$ can be written as

$$\hat{\bar{Y}}_d\left(\hat{\boldsymbol{\theta}}\right) = \frac{1}{N_d}\left(\sum_{j\in S_d} y_{dj} + \sum_{j\notin S_d} \hat{\mu}_{dj}^*\right), \tag{2.4}$$

where the second summation is taken over all nonsampled units with $\hat{\mu}_{dj}^*$ defined by

$$\hat{\mu}_{dj}^* = \boldsymbol{x}_{dj}^t\tilde{\boldsymbol{\beta}} + \gamma_d\left(\bar{y}_d - \bar{\boldsymbol{x}}_d^t\tilde{\boldsymbol{\beta}}\right). \tag{2.5}$$

2.2. Logistic model for binary responses

When y_{dj} is a binary response taking 1 or 0 according as the j^{th} unit has a certain attribute or not we consider a logistic model of the form

$$\log\left(\frac{\pi_{dj}}{1-\pi_{dj}}\right) = \boldsymbol{x}_{dj}^t\boldsymbol{\beta} + v_d, \tag{2.6}$$

where $\pi_{dj} = P(y_{dj} = 1|\,\boldsymbol{x}_{dj}, v_d)$. The π_{dj} can be expressed as

$$\pi_{dj} = \frac{\exp(\boldsymbol{x}_{dj}^t\boldsymbol{\beta} + v_d)}{1 + \exp(\boldsymbol{x}_{dj}^t\boldsymbol{\beta} + v_d)}. \tag{2.7}$$

The BLUP of (2.4), we call it BLUP type estimator for the small area proportion under the logistic mixed, can be written as

$$\hat{\bar{Y}}_d\left(\hat{\sigma}_v^2\right) = \frac{1}{N_d}\left(\sum_{j\in S_d} y_{dj} + \sum_{j\notin S_d} \hat{\pi}_{dj}^*\right), \tag{2.8}$$

where $\hat{\pi}_{dj}^*$ is obtained from (2.6) by substituting the estimators of $\boldsymbol{\beta}$ and v_d for the nonsampled units. We note that the $\hat{\bar{Y}}_d(\hat{\sigma}_v^2)$ is a composition of the mean of y_{dj} for the sampled units with the mean of predicted probabilities for the nonsampled units.

The coefficients $\boldsymbol{\beta}$ and the variance components σ_v^2 in (2.6) can be obtained by the ML or REML method by maximizing the marginal likelihood defined by

$$L\left(\boldsymbol{\beta}, \sigma_v^2|\,\boldsymbol{y}\right) = \prod_{d=1}^{D}\prod_{j=1}^{n_d}\int_{-\infty}^{\infty}\frac{\exp(y_{dj}(\boldsymbol{x}_{dj}^t\boldsymbol{\beta} + v_d))}{1+\exp(\boldsymbol{x}_{dj}^t\boldsymbol{\beta} + v_d)} \times \frac{e^{-\frac{v_d^2}{2\sigma_v^2}}}{(2\pi\sigma_v^2)^{\frac{1}{2}}}dv_d. \tag{2.9}$$

The integral with respect to the random effects distribution can not be solved analytically in general and we need numerical solutions. There are various numerical methods such as Gauss-Hermite quadrature method, Monte Carlo method, and penalized quasi-likelihood approximation. We adopt the Gauss-Hermite quadrature method to fit the logistic model (2.6) via Proc NLMIXED of SAS.

On the other hand González-Manteiga *et al.* (2007) applied Taylor series expansion to transform the logistic regression model (2.6) into the linear mixed model of the form

$$g(y_{dj}) \approx \boldsymbol{x}_{dj}^t\boldsymbol{\beta} + v_d + (y_{dj} - \pi_{dj})g'_{dj}(\pi_{dj}), \tag{2.10}$$

where $g_{dj}(\pi_{dj}) = \log\{\pi_{dj}/(1-\pi_{dj})\}$. The model (2.10) can be written in the form

$$\xi_{dj} = \boldsymbol{x}_{dj}^t\boldsymbol{\beta} + v_d + e_{dj} \tag{2.11}$$

with the assumption that e_{dj}'s are independent each other and also with the random effects v_d. Given v_d the variables ξ_{dj} are independent each other and the moments are $E(\xi_{dj}|v_d) = \eta_{dj}$ and $\mathrm{Var}(\xi_{dj}|v_d) = g'_{dj}(\pi_{dj})^2\sigma_{dj}$ with $\sigma_{dj} = \pi_{dj}(1-\pi_{dj})$. González-Manteiga *et al.* (2007) suggested the same estimator as given in (2.8) using the GLS estimator of $\boldsymbol{\beta}$ under the model (2.11).

3. Bootstrapping the Mean Squared Error

3.1. Estimation of mean squared error

When the $\hat{\mu}_d(\boldsymbol{\theta})$ in (2.3) is estimated by substituting $\hat{\boldsymbol{\theta}}$ Kackar and Harville (1984) used the following relationship under the normality of random effects and the translation-invariant property of $\hat{\boldsymbol{\theta}}$,

$$\mathrm{MSE}\left[\hat{\mu}_d(\hat{\boldsymbol{\theta}})\right] = \mathrm{MSE}\left[\hat{\mu}_d(\boldsymbol{\theta})\right] + E\left[\hat{\mu}_d(\hat{\boldsymbol{\theta}}) - \hat{\mu}_d(\boldsymbol{\theta})\right]^2. \tag{3.1}$$

According to Henderson (1975) without any distributional assumptions on v_d and e_{dj} the first component of (3.1) can be expressed in the form

$$\mathrm{MSE}\left[\hat{\mu}_d(\boldsymbol{\theta})\right] = (1-\gamma_d)\sigma_v^2 + \left(\bar{\mathbf{X}}_d - \gamma_d\bar{\boldsymbol{x}}_d\right)^t\left(\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\left(\bar{\mathbf{X}}_d - \gamma_d\bar{\boldsymbol{x}}_d\right). \tag{3.2}$$

But the second term $E[\mu_d(\hat{\boldsymbol{\theta}}) - \mu_d(\boldsymbol{\theta})]^2$ in (3.1) is generally not tractable except for the special case such as the balanced ANOVA model $y_{dj} = \mu + v_d + e_{dj}$ with $n_d \equiv r$. In estimating the finite population mean $\hat{\bar{Y}}_d$ the $\mathrm{MSE}[\hat{\bar{Y}}_d(\hat{\boldsymbol{\theta}})]$ approximately equals $\mathrm{MSE}[\hat{\mu}_d(\hat{\boldsymbol{\theta}})]$ provided $n_d/N_d \approx 0$ for every small area.

Because the MSE depends on the variance components we need to estimate the MSE. We may refer to Prasad and Rao (1990) for detailed discussion on the estimation of MSE and we simply introduce the formula by González-Manteiga *et al.* (2007) for the transformed model (2.11) with further notations. Let $\bar{\boldsymbol{x}}_{sd}^{\sigma} = \sum_{j\in S_d}\hat{\sigma}_{dj}\boldsymbol{x}_{dj}/\sigma_{sd.}$ and $\bar{\boldsymbol{x}}_{rd}^{\sigma} = \sum_{j\notin S_d}\hat{\sigma}_{dj}\boldsymbol{x}_{dj}/\sigma_{rd.}$, where $\sigma_{rd.} = \sum_{j\notin S_d}\hat{\sigma}_{dj}$ with $\hat{\sigma}_{dj} = \hat{\pi}_{dj}(1-\hat{\pi}_{dj})$. The variance matrix for the transformed variables ξ_{dj} in (2.11) is given by

$$\mathbf{V}_{\xi} = \sigma_v^2\mathbf{Z}\mathbf{Z}^t + \Sigma_e, \tag{3.3}$$

where $\mathbf{Z}$ is a block diagonal matrix with elements 1's and Σ_e is also a diagonal matrix having $g'(\pi_{dj})^2\sigma_{dj}$ as diagonals. The estimator of MSE by González-Manteiga *et al.* (2007) can be represented in terms of the following three components

$$\begin{aligned}
\hat{\mathbf{G}}_1 &= \hat{\sigma}_v^2\mathrm{diag}\left\{(1-\hat{\gamma}_d)\bar{\sigma}_{rd}^2\right\}\\
\hat{\mathbf{G}}_2 &= \left[\bar{\sigma}_{rd}\bar{\sigma}_{rd'}(\bar{\boldsymbol{x}}_{rd}^{\sigma} - \hat{\gamma}_d\bar{\boldsymbol{x}}_{sd}^{\sigma})(\mathbf{X}^t\boldsymbol{V}_{\xi}^{-1}\boldsymbol{X})^{-1}(\bar{\boldsymbol{x}}_{rd'}^{\sigma} - \hat{\gamma}_{d'}\bar{\boldsymbol{x}}_{sd'}^{\sigma})^t\right]_{D\times D}\\
\hat{\mathbf{G}}_3 &= 2\left\{\sum_{d=1}^{D}\sigma_{sd.}^2(1-\hat{\gamma}_d)^2\right\}\mathrm{diag}\left\{(1+\hat{\sigma}_v^2\sigma_{sd.})^{-3}\bar{\sigma}_{rd}^2\sigma_{sd.}\right\},
\end{aligned} \tag{3.4}$$

where $\bar{\sigma}_{rd} = \sigma_{rd.}/(N_d - n_d)$ and $\hat{\gamma}_d = \hat{\sigma}_v^2/(\hat{\sigma}_v^2 + \sigma_{sd.}^{-1})$ with $\sigma_{sd.} = \sum_{j\in S_d}\hat{\sigma}_{dj}$. The estimator of $\mathrm{MSE}(\hat{\bar{Y}}_d)$ by González-Manteiga *et al.* (2007), denoted by $\mathrm{mse}_d^{AP}(\hat{\bar{Y}}_d)$, is given by the d^{th} diagonal element of the matrix $\mathbf{G}_1 + \hat{\mathbf{G}}_2 + 2\hat{\mathbf{G}}_3$. That is

$$\mathrm{mse}_d^{AP}(\hat{\bar{Y}}_d) = \left[\mathrm{diag}\left(\hat{\mathbf{G}}_1 + \hat{\mathbf{G}}_2 + 2\hat{\mathbf{G}}_3\right)\right]_{(d)} \tag{3.5}$$

Table 3.1. Free throw data of NBA 15 top-scoring players

Player	n_d	y_d	p_d	$\hat{v}_d$	$\hat{\pi}_d$	$\hat{\bar{Y}}_d$	mse_d^B	mse_d^{AP}
Yao	13	10	0.769	0.0896	0.731	0.732	29.6193	86.9186
Frye	10	9	0.900	0.2481	0.761	0.763	44.8812	66.2024
Camby	15	10	0.667	−0.0786	0.696	0.695	27.9405	105.2871
Okur	14	9	0.643	−0.1139	0.689	0.687	31.8937	107.0592
Blount	6	4	0.667	−0.0401	0.704	0.704	29.0767	72.1310
Mihm	10	9	0.900	0.2481	0.761	0.763	44.5372	66.2024
Ilgauska	10	6	0.600	−0.1455	0.682	0.680	39.1234	100.3518
Brown	4	4	1.000	0.1794	0.748	0.750	36.3958	42.9510
Curry	11	6	0.545	−0.2303	0.663	0.661	44.5108	110.6431
Miller	10	9	0.900	0.2481	0.761	0.763	43.1030	66.2024
Haywood	8	4	0.500	−0.2317	0.663	0.660	49.3625	98.9308
Olowokan	9	8	0.889	0.2151	0.754	0.757	40.3247	65.9272
Mourning	9	7	0.778	0.0790	0.728	0.729	30.2221	77.3789
Wallace	8	5	0.625	−0.0960	0.692	0.691	37.1220	87.9868
Ostertag	6	1	0.167	−0.4705	0.608	0.602	115.6563	101.5380

where $[\text{diag (A)}]_{(d)}$ means the d^{th} diagonal element of the matrix A.

As an alternative method to the linear approximation the bootstrap method seems to be reasonable in the presence of small sample sizes. Instead of wild bootstrap by González-Manteiga *et al.* (2007) we propose a parametric bootstrap under the assumption of normality on the distribution of random effects. The parametric bootstrap algorithm consists of the following steps.

Step 1: Fit the logistic mixed model to obtain $\hat{\boldsymbol{\beta}}$ and $\hat{v}_d$, where $\hat{\boldsymbol{\beta}}$ denotes an ML estimator maximizing (2.9).

Step 2: Calculate every predicted value $\hat{\pi}_{dj}$ at $\boldsymbol{x}_{dj}$.

Step 3: Obtain the estimator $\hat{\bar{Y}}_d$ given in (2.8).

Step 4: Bootstrap the variables y_{dj}^* using $\hat{\pi}_{dj}$ from the Bernoulli distribution;

$$y_{dj}^* \sim B(1, \hat{\pi}_{dj}), \quad j = 1, \ldots, N_d, \quad d = 1, \ldots, D.$$

Step 5: Obtain $\bar{Y}_d^*$ and $\hat{\bar{Y}}_d^*$ in a similar way based on the bootstrapped $(y_{dj}^*, \boldsymbol{x}_{dj})$, where $\boldsymbol{x}_{dj}$ is the same as the population values.

Step 6: Repeat Step 4 and Step 5 B times and obtain the bootstrapped estimator

$$\text{mse}_d^B\left(\hat{\bar{Y}}_d\right) = B^{-1} \sum_{b=1}^{B} \left(\hat{\bar{Y}}_d^{*(b)} - \bar{Y}_d^{*(b)}\right)^2. \tag{3.6}$$

3.2. A practical example

We explain the proposed method through a practical example of NBA free throws success probabilities. The number of successes y_d among n_d trials for the fifteen center players in the NBA league of 2005–2006 is listed in Table 3.1, which is given in Agresti (2007) and also originally in nba.com. We assume the following mixed logistic model with no covariate term

$$\text{logit}(\pi_d) = \alpha + v_d, \tag{3.7}$$

where π_d are success probability for the d^{th} player, and v_d's are independent random variables having $N(0, \sigma_v^2)$.

In Table 3.1 p_d and $\hat{\pi}_d$ denote the sample proportion y_d/n_d and the predicted probability from the model (3.7), respectively. The $\hat{\bar{Y}}_d$ is the BLUP type composite estimator of (2.9) in which we artificially assumed the total number of trials to be $N_d = 500$ for the d^{th} player to formulate the small areas of finite population. The ML estimators are $\hat{\alpha} = 0.908$ and $\hat{\sigma}_v^2 = 0.1779$. The mse_d^{AP} is calculated from (3.5), and the bootstrap estimator mse_d^B is obtained from (3.6) with $B = 1{,}000$. We note that the sample proportions p_d varies from 0.167 to 1.0 with an extreme case of 0.608 for the player Ostertag. On the other hand the $\hat{\pi}_d$ based on the logistic mixed model varies between 0.61 and 0.76, and we note that the sample proportions shrink to the overall sample proportion $101/143 = 0.706$.

4. Monte Carlo Study

We design a simulation study to compare the efficiency of $\text{mse}_d^B(\hat{\bar{Y}}_d)$ and $\text{mse}_d^{AP}(\hat{\bar{Y}}_d)$ and also to asses the robustness of random effects distribution. We formulate a finite population consisting of $D = 20$ small areas of each size $N_d = 300$. Two types of sampling designs are considered; one is the case of equal sample sizes $n_d = 5, 10$ and the other the unequal sample sizes $n_d = 3$ or 7 according to the area number of odd or even. We consider the following logistic model with random intercepts

$$\log\left(\frac{\pi_{dj}}{1-\pi_{dj}}\right) = \alpha + \beta_1 x_{1dj} + \beta_2 x_{2dj} + v_d, \tag{4.1}$$

where x_{1dj} is taken to be uniform over $(-1, d/5)$ and x_{2dj} is Bernoulli with probability 0.5 and the coefficients are $\alpha = -1.25$, $\beta_1 = 0.5$, $\beta_2 = -2.75$. As a distribution for the random effects v_d we take $N(0, \sigma_v^2)$ and t-distribution with $3df$. The binary responses y_{dj}'s are generated from the Bernoulli distribution with probability π_{dj} given in (4.1). The number of Monte Carlo iterations and the bootstrap replications are $R = 200$ and $B = 200$, respectively.

To assess the efficiency of MSE estimators we define the d^{th} area MSE over R iterations by the quantity

$$\text{MSE}_d = \frac{1}{R}\sum_{i=1}^{R}\left(\hat{\bar{Y}}_{d(i)} - \bar{Y}_{d(i)}\right)^2, \tag{4.2}$$

where $\bar{Y}_{d(i)}$ is the d^{th} area mean at the i^{th} iteration of simulated population. With a little abuse of notation in this chapter we use the same notation $\bar{Y}_d$ to denote the average of $\bar{Y}_{d(i)}$ over R iterations, that is, $\bar{Y}_d = 1/R\sum_{i=1}^{R}\bar{Y}_{d(i)}$ with the corresponding estimator $\hat{\bar{Y}}_d$. We also define the mse_d as

$$\text{mse}_d = \frac{1}{R}\sum_{i=1}^{R}\text{mse}_{d(i)}, \tag{4.3}$$

where $\text{mse}_{d(i)}$ is the estimated MSE at the i^{th} iteration. The mse_d by the bootstrap and the approximation method will be denoted as mse_d^B and mse_d^{AP}, respectively. As an efficiency of mse_d by the bootstrap method we take the following quantity defined by

$$Q\left(\text{mse}_d^B\right) = \frac{1}{R}\sum_{i=1}^{R}\left(\text{mse}_{d(i)}^B - \text{MSE}_d\right)^2, \tag{4.4}$$

Table 4.1. Estimators of MSE for the equal sample sizes of $n_d = 5$ and 10

(a) when v_d's are random samples from $N(0, 4)$

Areas	n_d	$\bar{Y}_d$	$\hat{\bar{Y}}_d$	MSE_d	mse_d^B	mse_d^{AP}	Q_d^B	Q_d^{AP}
2	5	0.2806	0.3272	176.247	150.681	288.893	0.0037	0.0241
4	5	0.8243	0.7001	327.062	254.321	289.856	0.0199	0.0137
6	5	0.0309	0.1575	231.675	207.470	210.769	0.0097	0.0135
8	5	0.9101	0.7514	407.448	330.034	279.455	0.0296	0.0306
10	5	0.1823	0.2472	196.488	173.068	263.354	0.0046	0.0204
12	5	0.4342	0.4255	171.289	156.869	318.720	0.0035	0.0327
14	5	0.4730	0.4750	150.558	151.437	323.177	0.0037	0.0439
16	5	0.5089	0.4985	155.000	161.545	318.953	0.0034	0.0365
18	5	0.4984	0.5182	244.720	186.854	322.827	0.0072	0.0198
20	5	0.3585	0.3849	163.123	165.825	310.557	0.0052	0.0295
2	10	0.5234	0.5307	126.085	107.694	160.935	0.0008	0.0022
4	10	0.4145	0.4105	130.181	108.226	163.317	0.0008	0.0021
6	10	0.2159	0.2617	132.588	129.228	148.397	0.0008	0.0021
8	10	0.0336	0.0960	61.133	104.324	83.813	0.0042	0.0025
10	10	0.6613	0.6350	105.249	109.695	159.322	0.0004	0.0040
12	10	0.5447	0.5164	125.354	114.290	166.849	0.0006	0.0047
14	10	0.2169	0.2514	96.471	122.077	144.888	0.0019	0.0044
16	10	0.6401	0.6476	121.243	108.508	160.670	0.0006	0.0028
18	10	0.5412	0.5407	129.651	112.809	169.641	0.0007	0.0029
20	10	0.8410	0.8294	87.794	80.681	114.856	0.0009	0.0037

(b) when v_d's are random samples from t-distribution with $df = 3$

Areas	n_d	$\bar{Y}_d$	$\hat{\bar{Y}}_d$	MSE_d	mse_d^B	mse_d^{AP}	Q_d^B	Q_d^{AP}
2	5	0.3776	0.4074	110.139	101.189	233.653	0.0039	0.0210
4	5	0.4231	0.4438	115.169	109.046	245.985	0.0055	0.0231
6	5	0.5306	0.6038	163.943	110.713	228.601	0.0142	0.0173
8	5	0.6203	0.6703	117.131	93.957	210.340	0.0069	0.0148
10	5	0.7256	0.6495	156.660	93.230	219.112	0.0093	0.0109
12	5	0.5753	0.5683	77.561	88.427	234.079	0.0043	0.0309
14	5	0.4181	0.5048	189.195	104.333	236.068	0.0145	0.0104
16	5	0.7228	0.7586	108.544	97.497	199.015	0.0059	0.0180
18	5	0.7072	0.6726	89.358	78.360	219.189	0.0026	0.0252
20	5	0.8503	0.8733	117.570	81.317	130.879	0.0071	0.0134
2	10	0.5044	0.5088	47.376	47.374	127.519	0.0012	0.0075
4	10	0.4941	0.4710	58.529	56.849	133.743	0.0007	0.0067
6	10	0.6702	0.5992	116.077	73.374	125.540	0.0049	0.0011
8	10	0.2442	0.3720	243.112	113.173	126.525	0.0237	0.0147
10	10	0.5533	0.4970	111.211	71.122	130.409	0.0044	0.0014
12	10	0.4929	0.4900	55.209	49.555	125.888	0.0011	0.0060
14	10	0.3174	0.4484	252.004	117.182	134.755	0.0258	0.0151
16	10	0.5884	0.5357	99.242	63.562	128.658	0.0033	0.0019
18	10	0.6333	0.5779	107.166	76.516	141.956	0.0022	0.0021
20	10	0.4298	0.4585	65.722	50.808	125.747	0.0017	0.0046

where $\text{mse}_{d(i)}^B$ is the estimator of MSE by the bootstrap at the i^{th} iteration. Hereafter we simply denote the $Q(\text{mse}_d^B)$ as Q_d^B and we also use similar notation Q_d^{AP} for the mse_d^{AP}.

In the simulation results listed in Table 4.1 and Table 4.2 the MSE of (4.2) and its estimators are multiplied by 10^4, and both Q_d^B and Q_d^{AP} are multiplied by 10^2 for simplification of digits. As we

Table 4.2. Estimators of MSE for the unequal sample sizes

(a) when v_d's are random samples from $N(0,4)$

Areas	n_d	$\bar{Y}_d$	$\hat{\bar{Y}}_d$	MSE_d	mse_d^B	mse_d^{AP}	Q_d^B	Q_d^{AP}
1	3	0.9494	0.7660	448.324	254.074	403.162	0.0513	0.0471
2	7	0.5086	0.5286	136.542	133.482	224.228	0.0031	0.0106
3	3	0.5696	0.5857	227.358	198.445	450.142	0.0117	0.0794
4	7	0.7167	0.7043	125.271	109.015	203.832	0.0015	0.0097
5	3	0.0468	0.2911	726.203	447.054	396.824	0.1246	0.1289
6	7	0.5983	0.6031	154.925	122.772	216.559	0.0034	0.0068
7	3	0.6383	0.6177	200.922	184.687	450.915	0.0073	0.0921
8	7	0.8732	0.7982	150.216	169.855	172.123	0.0071	0.0064
9	3	0.9383	0.8110	244.299	152.690	347.365	0.0141	0.0722
10	7	0.9499	0.8643	123.776	125.985	148.477	0.0024	0.0107
11	3	0.5783	0.6151	186.343	165.385	439.700	0.0072	0.0914
12	7	0.6369	0.6377	137.473	120.718	220.655	0.0017	0.0104
13	3	0.5600	0.5926	208.339	182.313	446.342	0.0147	0.0821
14	7	0.4579	0.5430	244.569	203.853	229.474	0.0098	0.0039
15	3	0.5048	0.5216	222.425	209.325	478.461	0.0110	0.1018
16	7	0.1344	0.2562	283.561	304.354	215.687	0.0151	0.0120
17	3	0.4444	0.5452	320.298	280.682	474.253	0.0320	0.0468
18	7	0.8509	0.8258	117.221	105.895	163.787	0.0025	0.0123
19	3	0.7055	0.7404	309.204	227.540	363.077	0.0281	0.0444
20	7	0.7965	0.7492	144.613	161.367	201.919	0.0039	0.0087

(b) when v_d's are random samples from t-distribution with $df = 3$

Areas	n_d	$\bar{Y}_d$	$\hat{\bar{Y}}_d$	MSE_d	mse_d^B	mse_d^{AP}	Q_d^B	Q_d^{AP}
1	3	0.7080	0.5698	354.197	191.214	351.268	0.0555	0.0174
2	7	0.5670	0.5169	137.108	116.745	210.388	0.0050	0.0082
3	3	0.2983	0.3935	215.001	134.227	350.278	0.0161	0.0350
4	7	0.1864	0.2514	138.195	127.541	176.398	0.0039	0.0045
5	3	0.8181	0.6356	465.656	182.814	331.921	0.0982	0.0329
6	7	0.2544	0.3638	228.432	141.746	207.781	0.0149	0.0034
7	3	0.8278	0.6714	343.858	150.655	332.385	0.0455	0.0203
8	7	0.4278	0.5081	152.281	118.765	210.491	0.0050	0.0058
9	3	0.4816	0.4835	121.836	126.579	352.460	0.0102	0.0686
10	7	0.2750	0.4069	335.962	220.673	205.276	0.0284	0.0200
11	3	0.2533	0.3133	155.190	134.423	332.132	0.0072	0.0531
12	7	0.3669	0.3662	99.123	97.697	203.905	0.0018	0.0136
13	3	0.5530	0.5016	173.895	146.106	362.460	0.0137	0.0526
14	7	0.7267	0.6241	248.184	165.223	204.666	0.0160	0.0051
15	3	0.6116	0.5720	146.043	120.700	347.142	0.0068	0.0561
16	7	0.8521	0.7869	175.023	145.970	179.230	0.0037	0.0079
17	3	0.3864	0.4512	145.383	119.769	352.472	0.0092	0.0601
18	7	0.6439	0.5719	157.125	117.155	205.579	0.0056	0.0046
19	3	0.4768	0.5358	172.365	144.725	375.404	0.0140	0.0701
20	7	0.4151	0.4440	102.957	94.924	210.683	0.0027	0.0141

see in Table 4.1 the values of Q_d^B are smaller than those of Q_d^{AP} in general, and the values of MSE_d decrease when the sample size varies from $n_d = 5$ to $n_d = 10$. On the other hand the efficiencies of MSE estimators denoted by Q_d^B and Q_d^{AP} are improved when sample size becomes large. When the random effect has a t-distribution the MSE and its estimators are more variable across the

small areas. In this case the efficiencies of the estimators decrease compared to the case of normal random effects. The linear approximation method seems to be moderately good if normality of random effects is assumed and the sample size is $n_d = 10$. The parametric bootstrap is more robust to the violation from normality of the random effects distribution. The bootstrap has a tendency of little underestimation of the true MSE compared to the relatively large overestimation by the linear approximation method. The results of unequal sample sizes in Table 4.2 are similar to the case of equal sample sizes. We finally conclude that the parametric bootstrap is better than the linear approximation method in the sense of accuracy and robustness of estimation.

5. Conclusion and Further Comments

The estimation of small area proportions in the presence of small sample sizes is our main interest, where the auxiliary information of other areas are useful. In this paper we consider a logistic model with random effects to explain the random variation across small areas. The BLUP is popular in estimating the small area means under linear mixed model. We discuss a BLUP type estimator based on the logistic model with random intercepts to estimate small area proportions. The regression parameters are usually estimated by the numerical methods such as quasi Newton-Raphson based on the likelihood function. This type of estimator is the ML or REML estimator which can be obtained via the commonly used statistical packages.

The MSE is commonly used as an accuracy measure of an estimator but the MSE of BLUP type estimator depends on the unknown variance components. As an alternative to the approximation method for the estimation of MSE a parametric bootstrap is suggested. According to a Monte Carlo study the approximation method greatly overestimate the true MSE in contrast to a slight underestimation of the bootstrap method. The heavy computational burden of bootstrap method in performing a simulation study restricts the numbers of iterations and bootstrap replications to moderate sizes.

References

Agresti, A. (2007). *An Introduction to Categorical Data Analysis*, 2nd Ed., John Wiley & Sons, New York.

Ghosh, M. and Rao, J. N. K. (1994). Small area estimation: An appraisal, *Statistical Science*, **9**, 55–93.

González-Manteiga, W., Lombardia, M., Molina, I., Morales, D. and Santamaria, L. (2007). Estimation of the mean squared error of predictors of small area linear parameters under a logistic mixed model, *Computational Statistics & Data Analysis*, **51**, 2720–2733.

Hall, P. and Maiti, T. (2006). On parametric bootstrap methods for small area prediction, *Journal of the Royal Statistical Society, Series B*, **68**, 221–238.

Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model, *Biometrics*, **31**, 423–447.

Kackar, R. and Harville, D. A. (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models, *Journal of the American Statistical Association*, **79**, 853–862.

Lahiri, P. (2003). On the impact of bootstrap in survey sampling and small-area estimation, *Statistical Science*, **18**, 199–210.

Lohr, S. L. and Prasad, N. G. N. (2003). Small area estimation with auxiliary survey data, *The Canadian Journal of Statistics*, **31**, 383–396.

Prasad, N. G. N. and Rao, J. N. K. (1990). The estimation of the mean squared error of small-area estimators, *Journal of the American Statistical Association*, **85**, 163–171.

Rao, J. N. K. (2003). Practical issues in model-based small area estimation, In *Proceedings of Statistics Canada International Symposium 2003.*