Small Area Estimation in a Survey of High School Students in Iowa

Lu Lu[1], Michael D. Larsen [1]
Department of Statistics, Iowa State University, Ames, Iowa, 50011, U.S.A.[1]
Email: icyemma@iastate.edu, larsen@iastate.edu

Abstract

Iowa's State Board of Education (ISBE) conducted a stratified multi-stage sample survey to study the availability of employment preparation courses and the degree to which students in Iowa's public high schools enroll in those courses. The design stratifies districts in two dimensions, but given the budget and time constraints the sample of PSUs is small, which causes high variability in direct estimates. A hierarchical Bayesian (HB) analysis that borrows strength across strata with similar characteristics is suggested to improve efficiency and make better use of auxiliary information. Since the method is dependent on a valid model, effective model selection is crucial in HB estimation. The application of HB model selection and estimation for small areas is illustrated using a single simulated finite population based on the ISBE study.

KEY WORDS: Benchmarking; Generalized linear mixed models; Hierarchical Bayesian analysis; Model selection; Posterior predictive p-values.

1. Introduction

Small area estimation has received much attention in recent decades due to the increased need for accurate and reliable descriptions of small area characteristics for many public policy issues. Given the constraints of limited budgets and time the sample size in many surveys for educational and other studies is usually determined to produce accurate estimates at a relatively high level of aggregation, such as for states. As a result, there are often very small sample sizes allocated to individual small areas, such as school districts or substate educational areas. This will induce extremely unreliable direct estimates in these small areas, in which the policymakers are often interested as well.

Traditional indirect estimation methods produce more stable estimates in small areas by using synthetic or composite estimation. A synthetic estimator is an implicitly model-assisted estimator based on the assumption of small areas inheriting the same characteristics from the covering large area. It could dramatically reduce variances, but could cause "over-shrinkage" and potentially large bias in estimation due to an inappropriate implicit model assumption of homogeneity. The composite estimator, as a way of balancing the instability of a direct estimator and the potential bias of a synthetic estimator, utilizes both direct estimates at large areas and stabilized estimates at small areas. The exact way to balance the large and small area information needs to be specified.

Recent developments in small area estimation including empirical best linear unbiased prediction (EBLUP), empirical Bayes (EB) and hierarchical Bayes (HB) estimation have shown distinct advantages over traditional indirect estimators. Instead of using implicit models, these approaches utilize explicit models to delineate the dependent relationships among the local areas, especially allowing for modeling of local variation through complex error structures. More complex data structures such as geographic dependence, cross-sectional effects and time series correlation could be handled as well.

In 2004, representatives of Iowa's State Board of Education (ISBE) approached the Center for Survey Statistics and Methodology (CSSM) at Iowa State University (ISU) for help in planning a series of surveys. The purpose of one of the surveys is to study the availability of employment preparation (EP) courses and the degree to which students in Iowa's public high schools enroll in those courses. A primary concern of the survey is to assess the degree to which students in Iowa's public school districts, which vary greatly in size, community characteristics, and ruralness, have equal opportunities to prepare in school for employment, college, and life in general.

Due to the budget, time and policy restrictions a survey was conducted instead of a census. A stratified three-stage survey was designed to produce estimates of average numbers of EP courses of certain types taken by students for the State of Iowa and populations of small (less than 250 students in grades 9-12), medium (250 to less than 2,500 students) and large (2,500 or more students) school districts. Districts in Iowa are organized into twelve area education agencies (AEAs) for the purposes of administration and support. District size and AEA were used as stratifying variables. All large districts were included with certainty due to their extreme size. Medium and small districts were sampled with probability proportional to total enrollment size within stratum. For political reasons all schools in selected districts were included in data collection. A simple random sample of students was selected in each sampled school. The samples were split between grade nine and grade twelve students from general and special education groups.

The survey was designed to sample 60 schools and no more than 12,000 students. The 22 schools in eight large districts in seven AEAs were taken with certainty. The remaining 38 schools were split evenly between the medium and small school districts. From each of these size levels, 19 schools were selected from 12 strata. As a result, seven strata were assigned two PSUs and the remaining five strata that have relatively fewer districts had only one PSU sampled. Variance estimation for the strata with only one district sampled is a very challenging problem in surveys like ISBE's. In a one-per-stratum design, standard direct variance estimation is not applicable. Variance estimation based on collapsing strata is commonly used. But the approach produces a variance estimate only for a group of strata and the estimated variance itself tends to be highly unstable. In a preliminary study, we proposed to estimate variance in this one-per-stratum design using restricted generalized variance functions (RGVF), which produced better variance estimates in terms of a higher coverage rate for confidence intervals and more stable performance than a method utilizing collapsing (Lu and Larsen 2006).

Since the design takes a small sample of PSUs within strata, the direct estimator tends to produce highly unreliable estimates for individual strata. To make more efficient and reliable estimates of small area quantities, we consider using hierarchical Bayesian (HB) estimation. The method borrows strength across strata with similar characteristics and makes better use of auxiliary information than direct estimation. A fully Bayesian analysis provides a unified framework for surveys with small and large sample sizes and deals with nuisance parameters in a natural and appealing way. Monte Carlo integration techniques are employed to produce posterior estimates of parameters. A generalized linear mixed model (GLMM) is considered for small area modeling in Section 2. The HB estimator for the finite population mean under the GLMM is proposed in Section 3. The precision of the HB estimator is measured by its posterior variance. Three Bayesian methods of model comparison are considered for selecting an appropriate model. The performance of the estimators and model selection is illustrated using a single simulated finite population in Section 4. Section 5 contains a discussion about using HB estimation with careful model selection in small area estimation and suggests possible future research work.

2. Small Area Models

ISBE is interested in the characteristics of a multi-component population consisting of students from general and special education groups in ninth and twelfth grades. The population of twelfth grade students in Iowa's public high schools was chosen as a representative target population for the purpose of study. The inference for the multi-component population could be made by extending the univariate model to a multivariate model with an appropriate correlation structure.

Given the population structure and the sampling design, a GLMM is considered for modeling the population distribution. Let $y_{i,j,k,l}$ denote the number of EP courses taken by the l^{th} student from the k^{th} school in AEA j in size level i. Assume $y_{i,j,k,l}, l = 1, \cdots, n_{i,j,k}$, independently follow a Poisson distribution:

$$y_{i,j,k,l} | \lambda_{i,j,k} \sim \text{Poisson}\left(\lambda_{i,j,k}\right), \qquad (1)$$

where $\lambda_{i,j,k}$ is the rate of taking EP courses for students in the k^{th} school in AEA j in size level i. Then, we assume the rate of the Poisson distribution for each school is related to some auxiliary variables at the school level and random effects due to school size and AEA through a log-linear model

$$\log\left(\lambda_{i,j,k}\right) = x'_{i,j,k}\beta + \tau_i + \eta_j + \zeta_{i,j} + v_{i,j,k}. \qquad (2)$$

The $x_{i,j,k}$ of length p is a vector of covariate variables at the school level. The $\tau_i \sim N(0, \sigma_\tau^2)$, $\eta_j \sim N(0, \sigma_\eta^2)$ and $\zeta_{i,j} \sim N(0, \sigma_\zeta^2)$ are random effects from size, AEA, and the interaction between size and AEA. The random error term for the school is $v_{i,j,k} \sim N\left(0, \sigma_v^2\right)$. The model hyperparameters are β, σ_τ^2, σ_η^2, σ_ζ^2 and σ_v^2. If there was overdispersion in the Poisson distribution means, then we could consider a model like

$$\begin{array}{rcl} \lambda_{i,j,k} | \alpha, \mu_{i,j,k} & \sim & \text{Gamma}\left(\alpha, \alpha/\mu_{i,j,k}\right) \\ \log\left(\mu_{i,j,k}\right) & = & x'_{i,j,k}\beta + \tau_i + \eta_j + \zeta_{i,j}, \end{array} \qquad (3)$$

where α is a scale parameter that could be assumed common for the entire population (or verified across size levels or AEAs). The sample design is considered as ignorable because it is an inherent part of the model.

3. Hierarchical Bayes Analysis

In this section, we apply hierarchical Bayes (HB) analysis to the GLMM introduced in Section 2. Estimates of the posterior mean and variance of parameters are obtained from (MCMC) simulation.

3.1 Prior distributions

In a hierarchical Bayesian framework, we assume β, σ_τ^2, σ_η^2, σ_ζ^2, and σ_v^2 are mutually independent with diffuse prior distributions. Let β have a (locally) uniform distribution with $p(\beta) \propto 1$. Independently $\sigma_\tau^2 \sim \text{IG}\left(a_\tau, b_\tau\right)$, $\sigma_\eta^2 \sim \text{IG}\left(a_\eta, b_\eta\right)$, $\sigma_\zeta^2 \sim \text{IG}\left(a_\zeta, b_\zeta\right)$, and $\sigma_v^2 \sim \text{IG}\left(a_v, b_v\right)$, where IG denotes an inverse gamma distribution and a_τ, b_τ, a_η, b_η, a_ζ, b_ζ, a_v, and b_v are known positive constants. The constants usually are set to be very small to reflect our vague knowledge about the parameters. If a Poisson-Gamma model for overdispersion is employed, the scale parameter α can be assumed to have a prior distribution as $\alpha = w/(1-w)$, where $w \sim \text{Uniform}\,(0, 1)$.

3.2 Posterior estimation

Using a Gibbs sampler for computation, we independently simulate L parallel chains. After the convergence has been achieved for all parallel chains, a subsequence of

D iterates from each chain is retained for posterior estimation. The posterior mean and variance of $\lambda_{i,j,k}$ under the GLMM defined in (1) and (2) are given by

$$\begin{aligned} E\left(\lambda_{i,j,k}|y_s\right) &= E\{E(\lambda_{i,j,k}|\beta,\tau_i,\eta_j,\zeta_{i,j},\sigma_v^2,y_s)|y_s\} \\ &= E\{\exp(x'_{i,j,k}\beta+\tau_i+\eta_j+\zeta_{i,j}+\frac{1}{2}\sigma_v^2)|y_s\} \end{aligned}$$

and

$$\begin{aligned} & V\left(\lambda_{i,j,k}|y_s\right) \\ &= V\{E(\lambda_{i,j,k}|\beta,\tau_i,\eta_j,\zeta_{i,j},\sigma_v^2,y_s)|y_s\} + \\ &\qquad E\{V(\lambda_{i,j,k}|\beta,\tau_i,\eta_j,\zeta_{i,j},\sigma_v^2,y_s)|y_s\} \\ &= V\{\exp(x'_{i,j,k}\beta+\tau_i+\eta_j+\zeta_{i,j}+\frac{1}{2}\sigma_v^2)|y_s\} + \\ &\qquad E\{\exp[2(x'_{i,j,k}\beta+\tau_i+\eta_j+\zeta_{i,j})+\sigma_v^2] \\ &\qquad (e^{\sigma_v^2}-1)|y_s\} \\ &= E\{\exp[2(x'_{i,j,k}\beta+\tau_i+\eta_j+\zeta_{i,j}+\sigma_v^2)]|y_s\} - \\ &\qquad [E\{\exp(x'_{i,j,k}\beta+\tau_i+\eta_j+\zeta_{i,j}+\frac{1}{2}\sigma_v^2)|y_s\}]^2. \end{aligned}$$

These quantities can be estimated using the iterated simulates from MCMC as follows:

$$\begin{aligned} \hat{E}\left(\lambda_{i,j,k}|y_s\right) &= \frac{1}{LD}\sum_{l=1}^{L}\sum_{d=1}^{D}[\,\exp\{x'_{i,j,k}\beta^{(ld)}+ \\ &\qquad \tau_i^{(ld)}+\eta_j^{(ld)}+\zeta_{i,j}^{(ld)}+\frac{1}{2}\sigma_v^{(ld)^2}\}\,] \end{aligned} \tag{4}$$

and

$$\begin{aligned} \hat{V}\left(\lambda_{i,j,k}|y_s\right) &= \frac{1}{LD}\sum_{l=1}^{L}\sum_{d=1}^{D}[\,\exp\{2(x'_{i,j,k}\beta^{(ld)}+ \\ &\qquad \tau_i^{(ld)}+\eta_j^{(ld)}+\zeta_{i,j}^{(ld)}+\sigma_v^{(ld)^2})\}\,]-[\hat{E}\left(\lambda_{i,j,k}|y_s\right)]^2. \end{aligned} \tag{5}$$

The posterior covariance of $\lambda_{i,j,k}$ and $\lambda_{i',j',k'}$ is

$$\begin{aligned} & C\left(\lambda_{i,j,k},\lambda_{i',j',k'}|y_s\right) = \\ & C\{E(\lambda_{i,j,k}|\beta,\tau_i,\eta_j,\zeta_{i,j},\tau_{i'},\eta_{j'},\zeta_{i',j'},\sigma_v^2,y_s), \\ & E(\lambda_{i',j',k'}|\beta,\tau_i,\eta_j,\zeta_{i,j},\tau_{i'},\eta_{j'},\zeta_{i',j'},\sigma_v^2,y_s)|y_s\} + \\ & E\{C(\lambda_{i,j,k},\lambda_{i',j',k'}|\beta,\tau_i,\eta_j,\zeta_{i,j}, \\ & \tau_{i'},\eta_{j'},\zeta_{i',j'},\sigma_v^2,y_s)|y_s\} \\ &= C\{\exp(x'_{i,j,k}\beta+\tau_i+\eta_j+\zeta_{i,j}+\frac{1}{2}\sigma_v^2), \\ &\qquad \exp(x'_{i',j',k'}\beta+\tau_{i'}+\eta_{j'}+\zeta_{i',j'}+\frac{1}{2}\sigma_v^2)|y_s\}. \end{aligned}$$

It can be estimated by

$$\begin{aligned} & \hat{C}\left(\lambda_{i,j,k},\lambda_{i',j',k'}|y_s\right) = \\ & \frac{1}{LD}\sum_{l=1}^{L}\sum_{d=1}^{D}\exp\{(x_{i,j,k}+x_{i',j',k'})'\,\beta^{(ld)}+\tau_i^{(ld)}+ \\ & \tau_{i'}^{(ld)}+\eta_j^{(ld)}+\eta_{j'}^{(ld)}+\zeta_{i,j}^{(ld)}+\zeta_{i',j'}^{(ld)}+\sigma_v^{(ld)^2}\}- \\ & [\frac{1}{LD}\sum_{l=1}^{L}\sum_{d=1}^{D}\exp\{x'_{i,j,k}\beta^{(ld)}+\tau_i^{(ld)}+\eta_j^{(ld)}+ \\ & \zeta_{i,j}^{(ld)}+\frac{1}{2}\sigma_v^{(ld)^2}\}]\cdot[\frac{1}{LD}\sum_{l=1}^{L}\sum_{d=1}^{D}\exp\{x'_{i',j',k'}\beta^{(ld)}+ \\ & \tau_{i'}^{(ld)}+\eta_{j'}^{(ld)}+\zeta_{i',j'}^{(ld)}+\frac{1}{2}\sigma_v^{(ld)^2}\}], \end{aligned} \tag{6}$$

where the superscript (l,d) denotes the d^{th} iteration in the l^{th} chain in the retained subsequences.

Let $\theta_{i,j}$ denote the population mean for stratum (i,j), which is the quantity of interest. The $\theta_{i,j}$ could be considered as the sum of three terms

$$\begin{aligned} \theta_{i,j} &= N_{i,j}^{-1}\{\sum_{k\in s_{i,j}}\sum_{l\in s_{i,j,k}}Y_{i,j,k,l}+ \\ & \sum_{k\in s_{i,j}}\sum_{l\notin s_{i,j,k}}Y_{i,j,k,l}+\sum_{k\notin s_{i,j}}\sum_{l\in U_{i,j,k}}Y_{i,j,k,l}\}, \end{aligned} \tag{7}$$

where $N_{i,j}=\sum_{k\in U_{i,j}}N_{i,j,k}$ is the population size of stratum (i,j). A Bayesian estimate of $\theta_{i,j}$ is

$$\begin{aligned} E\left(\theta_{i,j}|y_s\right) &= N_{i,j}^{-1}\{\sum_{k\in s_{i,j}}n_{i,j,k}\bar{y}_{i,j,k}+ \\ & \sum_{k\in s_{i,j}}\left(N_{i,j,k}-n_{i,j,k}\right)E\left(\lambda_{i,j,k}|y_s\right)+ \\ & \sum_{k\notin s_{i,j}}N_{i,j,k}E\left(\lambda_{i,j,k}|y_s\right)\} \qquad (8) \\ &\equiv N_{i,j}^{-1}\{\sum_{k\in s_{i,j}}n_{i,j,k}\bar{y}_{i,j,k}+l'_{i,j}E\left(\lambda|y_s\right)\}. \end{aligned}$$

In the above, $\lambda=\{\lambda_{i,j,k}\}$ is a parameter vector of rates of Poisson distributions for schools in the entire population and $l'_{i,j}=\{0,\cdots,0,\tilde{l}_{i,j},0,\cdots,0\}$ is the vector of coefficients for stratum (i,j). In the latter expression, $\tilde{l}_{i,j}=\{l_{i,j,k}\}_{k\in U_{i,j}}$ is the vector of values $l_{i,j,k}$ in stratum (i,j), where $l_{i,j,k}=(N_{i,j,k}-n_{i,j,k})$ if $k\in s_{i,j}$ and $N_{i,j,k}$ if $k\notin s_{i,j}$. The set $s_{i,j}$ denotes the sample.

The HB estimator of $\theta_{i,j}$ is proposed as

$$\hat{\theta}_{i,j}=N_{i,j}^{-1}\{\sum_{k\in s_{i,j}}n_{i,j,k}\bar{y}_{i,j,k}+l'_{i,j}\hat{E}\left(\lambda|y_s\right)\}. \tag{9}$$

The posterior variance of $\hat{\theta}_{i,j}$ is

$$V\left(\theta_{i,j}|y_s\right)=N_{i,j}^{-2}\{l'_{i,j}V\left(\lambda|y_s\right)l_{i,j}\} \tag{10}$$

which can be estimated by plugging $\hat{V}\left(\lambda|y_s\right)$ into (10), where the diagonal and off-diagonal elements of $\hat{V}\left(\lambda|y_s\right)$ are calculated by (5) and (6), respectively.

3.3 Benchmarked HB estimators

In many surveys, even though we have very small sample sizes in small areas, we usually still have enough sample in a larger region consisting of a group of small areas to produce a reliable estimate for the large region. Assume that an accurate and reliable direct estimate for an aggregation of small areas is available. We want to benchmark the HB estimators for individual areas such that the aggregation of the benchmarked HB (BHB) estimates equals the direct estimate over the larger region (You, Rao, and Dick 2004).

In the EP survey, we have relatively reliable direct estimates at size levels. The benchmark property with respect to the size level direct estimate is

$$\sum_j N_{i,j}\hat{\theta}_{i,j}^{BHB} = \sum_j N_{i,j}\hat{\bar{y}}_{i,j}, \qquad (11)$$

where $i \in \{\text{size level: } 1 = \text{large}; 2 = \text{medium}; 3 = \text{small}\}$, $j \in \{12 \text{ AEAs}\}$, and $\hat{\bar{y}}_{i,j}$ denotes the direct estimate of the population mean for stratum (i,j). The raking-benchmarked HB (RBHB) estimator for stratum (i,j) can be obtained as

$$\hat{\theta}_{i,j}^{RBHB} = \hat{\theta}_{i,j}^{HB}\frac{\sum_j N_{i,j}\hat{\bar{y}}_{i,j}}{\sum_j N_{i,j}\hat{\theta}_{i,j}^{HB}}. \qquad (12)$$

The posterior mean square error (PMSE) of the BHB estimator is

$$PMSE\left(\hat{\theta}_{i,j}^{BHB}\right) = V\left(\theta_{i,j}|y_s\right) + \left(\hat{\theta}_{i,j}^{BHB} - \hat{\theta}_{i,j}^{HB}\right)^2 \qquad (13)$$

(You, Rao, and Dick 2004). As long as practically feasible, we can benchmark to two or more levels of standards. The benchmarked HB estimator is design consistent in the larger region, which is an attractive property. Due to benchmarking the BHB estimator should be more robust to model failure than the HB estimator. When the model is misspecified, benchmarking could correct the bias of the HB estimator to some degree. The PMSE derived under the model, however, could be seriously inflated due to a large bias correction. Therefore, effective model selection and model checking are highly important.

3.4 Model selection

Model assessment or model comparison has always been an important dimension of model-based inference. If a statistical model is not appropriate for a given relationship in the population, then analysis based on the model could be very misleading. The appropriateness of a model is measured by not only the form of model structure but also the involvement of covariate information. Variable selection concerns which of the possibly several predictor variables to use in a model. The problem of variable selection can be viewed essentially as a problem of model selection in a statistical application.

Traditional procedures of model comparison and variable selection rely on Bayes factors. To use Bayes factors, it is necessary to specify proper prior distributions for the parameters and models. This can be heavy work to specify prior distributions for all models under consideration, especially if there is a large number of potential covariate variables available. In addition, the posterior model probabilities are generally sensitive to the choice of prior parameters, which in general is not desirable.

Alternatively, recent developments have been focussed on a predictive approach, which is applicable for utilizing improper prior distributions as long as the resulting posterior distributions are proper. The method can be used not only for the comparison between nested models but also for the comparison across a large class of plausible non-nested models. We will discuss and apply three Bayesian predictive methods for model comparison.

The first method is based on the posterior predictive p-value, which measures the probability that the predictive data could be more extreme than the observed data in terms of a certain "discrepancy" measure. The "discrepancy" measure could be some test statistic or more generally could involve the unknown "nuisance" parameters as well. One of the commonly used "discrepancy" measures is the χ^2 discrepancy defined as $\mathrm{X}^2(y;\theta) = \sum_{i=1}^n \frac{(y_i - E(y_i|\theta))^2}{Var(y_i|\theta)}$, where $y = (y_1, \cdots, y_n)$ is a vector of independent observations and θ is a vector of parameters. According to Gelman, Meng and Stern (1996), the posterior predictive p-value can be approximated by the frequency of the predictive discrepancy (based on replicated predictive values) exceeding the realized discrepancy (based on observed data) among a large number of posterior predictions.

The second approach is the L-criterion proposed by Laud and Ibrahim (1995). The method measures the performance of a model by evaluating expected posterior predictive errors. The measurement is implemented through an imaginary device of a replicate experiment which is assumed to be done under the same conditions as the current experiment. For a given model m, define $L_m^2 = E\{(Z-y)'(Z-y)\} = \sum_{i=1}^n[\{E(Z_i) - y_i\}^2 + Var(Z_i)]$, where Z denotes the vector of the response values in the replicated experiment. The expectation is taken with respect to the predictive density of a replicated experiment (PDRE) defined as $p(z|y,m) = \int p\left(z|m,\theta^{(m)}\right)\pi\left(\theta^{(m)}|m,y\right)d\theta^{(m)}$. The density $\pi\left(\theta^{(m)}|m,y\right) \propto \pi\left(\theta^{(m)}|m\right)p\left(y|m,\theta^{(m)}\right)$ is the posterior distribution for $\theta^{(m)}$ under model m given observed data $Y = y$. The L_m^2 could be considered as a measure of how close the predictive data is to the observed data accounting for the variability of the predictions. Small values of L_m^2 indicate good models. Laud and Ibrahim (1995) referred to $L_m = \sqrt{L_m^2}$ as the L-criterion due to its convenience of use as a measure in the same scale as the response variable. They also suggested to quantify the uncertainty that is inherent in the criterion values by calculating the standard deviation of the criterion with respect to the marginal distribution of the outcome variable. The calibration number for the L-criterion is defined as $S_{L_{m^*}} = [Var\{L_{m^*}(Y)\}]^{1/2}$, where m^* denotes the model with the smallest criterion value. Hoeting and Ibrahim (1998) defined a comparison score as $\phi_m = \frac{L_m - L_{m^*}}{S_{L_{m^*}}}$, which measures the number of calibration units that a given model is from the model with the smallest criterion value. A simple model with a relatively small comparison score, say less than 2, is preferred.

The third method is based on the deviance information criterion (DIC). The deviance defined as $D(y, \theta) = -2 \log p(y)$ has an important role in statistical model comparison due to its connection to the Kullback-Leibler (KL) information measure. The expected deviance as a measure of predictive accuracy is therefore often used as a measure of overall model fit. The estimate of expected posterior deviance is given by $\hat{D}_{avg}(y) = \frac{1}{L}\sum_{l=1}^{L} D\left(y, \theta^{l}\right)$, where l denotes the number of iteration. The model complexity is measured by the effective number of parameters of a Bayesian model, which could be approximated by $p_D = \hat{D}_{avg}(y) - D_{\hat{\theta}}(y)$, where $D_{\hat{\theta}}(y)$ is the deviance at a point estimate of θ such as the posterior mean. The DIC is defined as the sum of the expected posterior deviance and the effective number of parameters, which could be considered as a Bayesian measure of fit or adequacy, penalized by an additional model complexity term p_D. Spiegelhalter et al. (2002) show that in models with negligible prior information DIC will be approximately equivalent to Akaike's criterion (AIC).

4. Illustration

To illustrate the performance of the proposed estimators and model comparison methods, we simulated a single finite population of EP courses taken by twelfth grade students from Iowa's public high schools from a Poisson log-linear model with random effects from size levels and AEAs. Population sizes in the simulation match actual population sizes in Iowa's school districts in 2004. One sample data set was drawn from the simulated population under the stratified three-stage design.

Seven models consisting of different combinations of auxiliary variables and random effects are considered:

$$
\begin{aligned}
&\text{Model 1:} && \log(\lambda_{i,j,k}) = b_0 + b_1 x_{i,j,k;1} \\
&\text{Model 2:} && \log(\lambda_{i,j,k}) = b_0 + b_1 x_{i,j,k;1} + \tau_i \\
&\text{Model 3:} && \log(\lambda_{i,j,k}) = b_0 + b_1 x_{i,j,k;1} + \tau_i + \eta_j \\
&\text{Model 4:} && \log(\lambda_{i,j,k}) = b_0 + b_1 x_{i,j,k;1} + \tau_i + \eta_j + \zeta_{i,j} \\
&\text{Model 5:} && \log(\lambda_{i,j,k}) = b_0 + \tau_i + \eta_j \\
&\text{Model 6:} && \log(\lambda_{i,j,k}) = b_0 + \sum_{q=1}^{5} b_q x_{i,j,k;q} \\
&\text{Model 7:} && \log(\lambda_{i,j,k}) = b_0 + b_1 x_{i,j,k;1} + \tau_i + \eta_j + v_{i,j,k}
\end{aligned}
$$

The covariate variables $x_{i,j,k;q}, q = 1, \cdots, 5$ correspond to auxiliary information about the total enrollment size, the amount of funding per student, and the percentage of male students, white students, and students having free or reduced price lunch. The variables have been transformed using logarithmic or power transformations to produce more uniform or symmetric distributional shapes. The parameter b_0 is the intercept and $b_q, q = 1, \cdots, 5$ are the regression coefficients. Among these models, model 3 is the model from which the population was simulated. Model 3 and any model from the rest of models except model 6 are nested. Model 6 is a GLM involving all five covariate variables but no random effects.

The prior distributions for model parameters are specified in Section 3.1. For each model, by using the Gibbs sampling algorithm, we independently simulated $L = 3$ parallel Markov chains, each of length $10,000$ iterations. The first 5,000 iterations for each chain are deleted as a "burn-in" period. By thinning to every 5^{th} iteration, 1,000 iterates from each chain are retained for posterior estimation.

As an example of Bayesian model selection, Table 1 shows the results of comparing seven models using the three methods discussed in Section 3.4. According to the posterior predictive p-values (PPPs), models 1, 2, 5 and 6 show strong evidence of model failure due to the extreme pattern of observed data relative to the posterior predictive data based on the assumed model. Models 3, 4 and 7 have no indication of model inadequacy. By choosing the most parsimonious model among models with unextreme PPP, the true model (model 3) will be selected. When using the L-criterion, model 4 has the smallest L_m value. Calibrated by the standard deviation of the criterion value under model 4, the comparison scores (CSs) for models 1, 2, and 6 are larger than a value of 4, which is too extreme in terms of calibration of inherent variation of criterion values. Model 5 is the smallest model with a not extreme CS. Among the models have the smallest DIC value, model 3 will be selected due to its simplest form. The PPP and DIC criteria successfully choose the true model. The L-criterion selects model 5 which is only different from the true model by omitting the first covariate variable x_1. The reason could be the coefficient of x_1 is very small and the range of x_1 is also very short so that the effect of the first covariate term is small relative to other effects.

By comparing the HB estimates under models 3 and 5, model 5 produces larger absolute relative bias (ARB) and higher posterior mean square error (PMSE) in most of strata. The HB estimator derived under model 5 still shows advantages over the direct ratio estimator. Ba-

Table 1: Bayesian model selection criteria in the example. Model 3 was used to generate the data. *PPP*=posterior predictive p-value. *CS*=comparison score for the L_m statistic. *DIC*=deviance information criterion. p_D=effective number of parameters. Model 3 is smaller (p_D) than models 4 & 7.

	PPP	L_m	*CS*	*DIC*	p_D
M1	0.000	249.00	6.300	20410	1.96
M2	0.000	245.28	4.517	20160	3.91
M3	**0.136**	235.92	**0.041**	**19500**	14.74
M4	**0.143**	235.84	**0.000**	**19500**	20.81
M5	0.014	237.47	0.779	19610	13.80
M6	0.000	248.13	5.885	20350	4.05
M7	**0.146**	235.86	**0.010**	**19500**	24.58

sically, these Bayesian model comparison methods work well in selecting an appropriate model for further analysis. Referring to more than one criteria if practically feasible should be helpful in making a good decision.

In our preliminary study using direct esimation, we chose the ratio estimator because it produced better estimates than the Horvitz-Thompson estimator in terms of smaller variance and mean square error (MSE) in the Monte Carlo study (Lu and Larsen 2006). Now we compare the model-based HB estimator with the design-based ratio estimator based on the absolute relative bias (ARB) and the square root of mean square error (RMSE) for individual strata. The ARB is defined as the absolute value of the relative bias of the estimate over the realized finite population value. The MSE of the ratio estimator is estimated through Monte Carlo simulation. The posterior MSE of the HB estimator equals to the posterior variance under the model. The PMSE of the BHB estimator was calculated by (13) in Section 3.3.

To see the performance of different estimators at the small area (stratum) level, we choose the twelve strata consisting of medium districts as a representative of presenting estimation results.

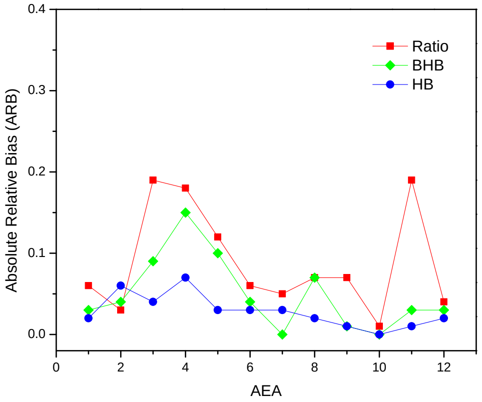

Figure 1: The Absolute Relative Bias (ARB) of ratio, HB, and BHB estimates under the true model (model 3) based on a single simulated finite population.

Figure 1 shows the absolute relative bias (ARB) of ratio, HB and BHB estimates over the realized (true) finite population mean for individual strata, when the true (correct) model is employed. The strata are sorted by the population size of PSUs. Larger strata get more PSUs in the sample. The five strata on the left have one PSU sampled and the seven strata on the right have two PSUs sampled. For the single randomly selected sample, the ratio estimator produces consistently larger ARBs for all except one stratum. Three out of twelve strata have ARBs of ratio estimates almost as high as 15 − 20% of the realized finite population mean. The ARBs for HB estimates are less than 8% for all medium strata and less than 4% for larger medium strata with two PSUs sampled. The ratio estimator shows much higher variation than the HB estimator at the small area level when the model is correct. As a hybrid of ratio and HB estimators, the BHB has ARBs and variation of estimates in between.

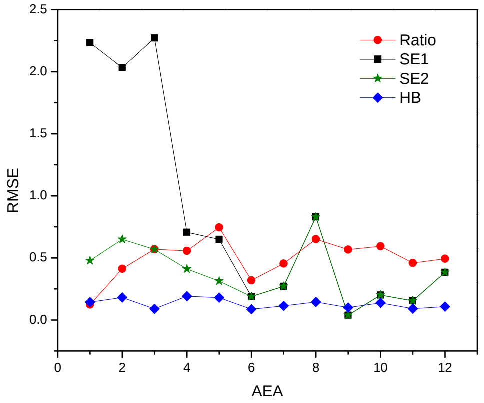

Figure 2: The RMSE (ratio) and SE (SE1, SE2) of the ratio estimate and RPMSE (HB) of the HB estimate under the true model (model 3). The RMSE of ratio estimate is obtained based on 1,000 simulated samples. The SEs of the ratio estimate and the RPMSE of the HB estimate is based on a single simulated finite population.

Figure 2 displays the root of mean square error (RMSE) of ratio and HB estimators. The MSE of the ratio estimator is estimated through 1,000 replicated simulations of the sample from the finite population. The posterior MSE of the HB estimator is derived under the true model. Since in reality we usually have only one set of sample data, it is difficult to estimate MSE through replicated samples that are really generated from the finite population. People usually use the standard error to quantify the design variation of direct estimator. Unfortunately, in a one-PSU-per-stratum design, there are not enough degrees of freedom to estimate variance directly. Besides the concern of reliability of the direct estimator, the assessment of precision of the estimator is also a challenging problem. Figure 2 also shows two kinds of standard errors (SEs) of the ratio estimator for strata with one PSU sampled. The SE1 was obtained by collapsing strata followed by synthetic variance redistribution. The SE2 was estimated by using the restricted generalized variance function method (Lu and Larsen 2006). In the case of our application, the collapsing strata estimator significantly overestimated the variances in small areas. The generalized variance function method did better, but since it is still design-based in substance, it would inherent the instability of the direct estimator in small sample cases. In contrast with the direct estimator, the HB estimator with a properly specified model produces more reliable estimates in terms of smaller PMSE. The advantage of using a model-based estimator is significant in terms of producing more efficient and reliable estimates. Additionally, the HB method addresses analytical inference in a unified framework for surveys with small and large sample sizes and deals with the nuisance parameters in a natural way, thereby simplifying the production of appropriate variance estimates in small sample cases. HB shows its great advantage in this regard compared to not only the direct estimator but also other model-based estimators such as EBLUP and EB.

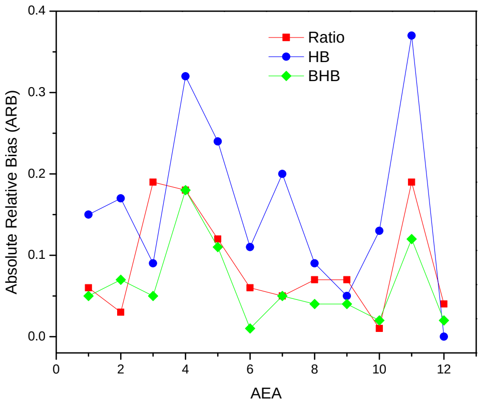

Figure 3: The Absolute Relative Bias (ARB) of ratio, HB, and BHB estimates under an inadequate model (model 2) based on a single simulated finite population.

However, just like all model-based estimators, the HB estimator is also vulnerable to mis-specification of the model. Figure 3 shows the ARB of ratio, HB and BHB estimates when an inadequate model (model 2) is used. The HB estimator derived based on the smaller model, which fails to address the random effect from AEAs and had shown strong evidence of model inadequacy in the previous model checking, produces significantly larger bias than the ratio estimator for most strata. Four strata have ARBs more than 20%. By benchmarking to the direct estimates at size and AEA levels, the BHB estimator successfully ”corrects” the bias induced by model mis-specification and produces comparable ARBs with the ratio estimator. On the other hand, BHB pays a price for ”correcting” the bias: the PMSE is inflated.

When the model is tolerable, the inflation of PMSE for the BHB estimator might not be too bad. But if the model is very poorly specified, the PMSE for BHB could be extremely large. Figures 4 and 5 display the RPMSEs of HB and BHB estimators under the true model and the smaller model respectively. The HB estimator has very small RPMSEs in both cases. The BHB estimator always has larger RPMSE due to the ”correction” of bias. The inflation of PMSE using the smaller model is much bigger than using the true model. This is because model failure caused a serious bias of the HB estimator and a corresponding big bias correction term for BHB. Therefore, serious inflation of PMSE of BHB relative to the HB estimator could be an indication of a poorly specified model.

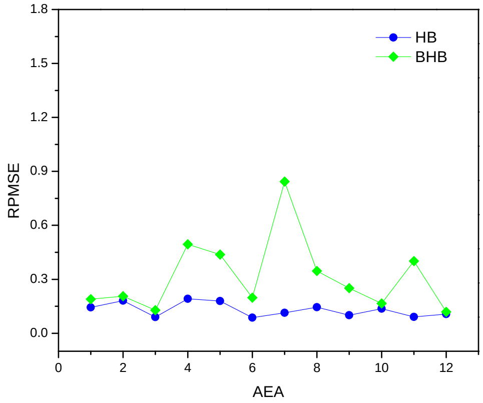

Figure 4: The root of posterior mean square error (RPMSE) of HB and BHB estimates under the true model (model 3) based on a single simulated finite population.

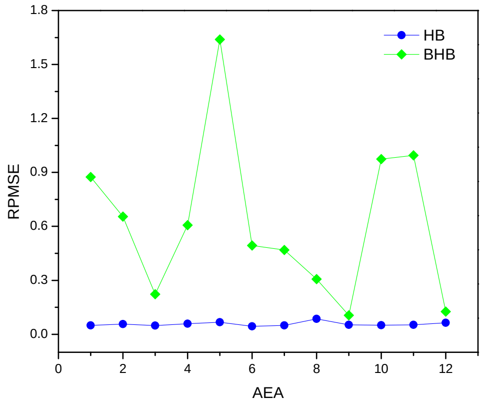

Figure 5: The root of posterior mean square error (RPMSE) of HB and BHB estimates under an inadequate model (model 2) based on a single simulated finite population.

From the above figures, we see great advantages of using a model-based estimator in terms of producing efficient and reliable estimates when a model is properly chosen, especially for problems of inference with small sample sizes. We also see the issue of the HB estimator being vulnerable to model mis-specification, which could cause serious estimation bias. By benchmarking to reliable direct estimates at high levels of aggregation of small areas, the BHB estimator could ”correct” the bias in small areas to some degree and achieve some nice properties, such as design-consistency and reliable estimates at a larger region. The disadvantage is that the PMSE could be dramatically inflated if the model is poorly specified. Therefore, careful model specification is crucial in model-based estimation.

5. Summary And Discussion

A survey on transcripts of Iowa's public high school students motivated an examination of small area estimation through model-based inference. The method of producing more reliable estimates for areas with small sample sizes than direct estimation were studied from a full Bayesian perspective.

The hierarchical Bayes (HB) approach was used to obtain the posterior estimates of the average number of EP courses taken by twelfth grade high school students for strata defined by district size and AEA and populations of aggregations of strata. When an appropriate model is used, the HB estimator outperforms the ratio estimator by borrowing strength across strata in terms of producing consistently smaller absolute relative bias (ARB) and root of mean square error (RMSE) for individual strata.

The use of the HB method could be very helpful in gaining more efficiency in estimation. It could, however, produce a misleading survey inference if the model is poorly specified. Effective model selection is crucial in the HB analysis. The issue of model selection not only includes selecting proper model structure but also includes selecting covariate variables and proper forms of transformations of the variables. In the illustration, we examined three Bayesian model comparison methods. The posterior predictive p-value and deviance information criteria successfully chose the true model. The L-criterion selected a model close to the true model, which still produces HB estimates better than the direct estimates. We are not suggesting these three are better than other methods for Bayesian model selection. These are just three methods that are easy to use and usually do a nice job. There are many other methods that we have not even mentioned here. The point is an effective model selection is a good start and also a crucial basis for making an efficient survey inference using the HB approach. With a carefully chosen model, the HB estimator should outperform the direct estimator in small area estimation.

Future study will pay more attention to examining effective methods of choosing proper forms of transformations of predictive variables and developing an efficient strategy to combine the selection of variables and transformations in the application of Bayesian model selection. In large-scale surveys, since it is practically inefficient or impossible to compare all possible models with various combinations of variables, we also hope to explore some more efficient methods to choose promising models such that the scope of model selection could be narrowed down to allow us to compare the models one-by-one based on the criterion-based methods. Further, for predictive survey inference, instead of choosing one single model and assuming the model is the true model to carry on the analysis, we could count in the model uncertainty by averaging the predicted values over a group of promising models weighted by the posterior probabilities of the models. Therefore, besides Bayesian model selection, Bayesian model averaging might be another option for future study.

Acknowledgments

The authors wish to thank Iowa's State Board of Education for partial support of this work. This research was also supported by a grant from the American Educational Research Association which receives funds for its "AERA Grants Program" from the National Science Foundation and the National Center for Education Statistics of the Institute of Education Sciences (U.S. Department of Education) under NSF Grant #REC-0634035. Opinions reflect those of the authors and do not necessarily reflect those of the granting agencies.

References

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (1995), *Bayesian Data Analysis*, London: Chapman and Hall.

Gelman, A., Meng, X.L., and Stern, H. (1996), "Posterior predictive assessment of model fitness via realized discrepancies", *Statistica Sinica*, **6**, 733-807.

Hoeting, J. and Ibrahim, J.G. (1998), "Bayesian predictive simultaneous variable and transformation selection in the linear model", *Journal of Computational Statistics and Data Analysis*, **28**, 87-103.

Laud, P.W. and Ibrahim, J.G. (1995), "Predictive model selection", *Journal of the Royal Statistical Society*, Series B, **57**, 247-262.

Lu, L., and Larsen, M.D. (2006), "A comparison of methods for a survey of high school students in Iowa", *Proceedings of the Survey Research Methods Section*, American Statistical Association.

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

Rao J.N.K. 2003, *Small Area Estimation*, New York: Wiley.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., and Linde A. (2002), "Bayesian measures of model complexity and fit", *Journal of the Royal Statistical Society*, Series B, **64**, 583-639.

You, Y., Rao, J.N.K., and Dick, P. (2004), "Benchmarking hierarchical Bayes small area estimators in the Canadian census undercoverage estimation", *Statistics In Transition*, **6**, 631-640.