Hierarchical Model Selection Using a Benchmark Discrepancy

Lu Lu[*] Michael D. Larsen[†]

Abstract

In the context of small area estimation, hierarchical Bayesian (HB) models are often proposed to produce more reliable estimators of small area quantities than direct estimates, such as design-based survey estimators. A method that benchmarks HB estimates with respect to higher level direct estimates and measures the relative inflation in the posterior mean square error of distributions due to benchmarking is developed to evaluate the performance of hierarchical models. The benchmarked hierarchical Bayesian posterior predictive model comparison method is shown to be able to select proper models effectively in a simulation study. The method is then applied to fitting models to a stratified multi-stage sample survey conducted by Iowa's State Board of Education. In this study a small sample of school districts was selected from a two-way stratification of school districts. The survey strata serve as small areas for which hierarchical Bayesian estimators are suggested. Here the method is used to select a generalized linear mixed model for the survey data. Potential applications extend beyond the survey and education contexts.

Key Words: generalized linear mixed models; Poisson-gamma model; Poisson-lognormal model; posterior predictive checks; small area estimation.

1. Introduction

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. Budget, time and policy restrictions influenced the survey design. 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 , medium and large 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. As a result of restricted sample size of schools, for each of the medium and small size levels, seven strata were assigned two PSUs and the remaining five strata that have relatively fewer districts had only one PSU sampled.

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.

Generalized linear mixed models (GLMMs) are considered for small area modeling in Section 2. The HB estimators for the finite population mean under the GLMMs are proposed in Section 3. The precision of the HB estimators is measured by their posterior variances. In Section 4, a new discrepancy measure based on evaluating the inflation of posterior mean square error due to benchmarking the HB estimates with respect to the reliable direct estimates in the larger regions is developed. The use of the new discrepancy measure in posterior predictive checking also is defined. In Section 5, the performance of the estimators and the model selection methods is examined using a single simulated finite population. The methods are applied for analyzing the actual data from the ISBE survey in Section 6. Section 7 contains a discussion about using HB estimation with careful and efficient model selection in small area estimation and suggests possible future research work.

[*] Department of Statistics, Iowa State University, Ames, Iowa, 50011, U.S.A. Email: icyemma@iastate.edu

[†] Department of Statistics, Iowa State University, Ames, Iowa, 50011, U.S.A. Email: larsen@iastate.edu

2. Small Area Models

The survey for ISBE recorded the number of EP courses taken by students in a sample of students from Iowa's public high schools. School districts are grouped in to AEAs and are categorized by size. Some districts have multiple high schools. Other applications in a wide array of areas use comparable designs: stratification (or blocking) factors and clustered (nested) units within cells.

Given the population structure and the sampling design, two GLMMs are considered for modeling the population distribution. In both models, let y_{ijkl} 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 $y_{ijkl}, l = 1, \cdots, n_{ijk}$, independently follow a Poisson distribution: $y_{ijkl}|\lambda_{ijk} \sim \text{Poisson}\left(\omega_{ijkl}\lambda_{ijk}\right)$, where λ_{ijk} is the rate of taking EP courses per semester for students in the k^{th} school in AEA j in size level i and ω_{ijkl} is the number of semesters that the l^{th} student has had in the school.

In the *Poisson-Lognormal* model, 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 district size and AEA through a Lognormal model: $\log\left(\lambda_{ijk}\right) = x'_{ijk}\beta + \tau_i + \eta_j + \zeta_{ij} + v_{ijk}$. The x_{ijk} of length p is a vector of covariate variables at the school level. The $\tau_i \sim N(0, \sigma^2_\tau)$, $\eta_j \sim N(0, \sigma^2_\eta)$ and $\zeta_{ij} \sim N(0, \sigma^2_\zeta)$ are independent random effects from size, AEA, and the interaction between size and AEA. The random error term for the school is $v_{ijk} \sim N\left(0, \sigma^2_v\right)$. The model hyperparameters are β, σ^2_τ, σ^2_η, σ^2_ζ and σ^2_v.

In the *Poisson-Gamma* model, the Poisson rate is assumed to follow a Gamma distribution with a mean related to the random effects and auxiliary variables through a log-linear model: $\lambda_{ijk}|\alpha, \gamma_{ijk} \sim \text{Gamma}\left(\alpha, \alpha/\gamma_{ijk}\right)$ and $\log\left(\gamma_{ijk}\right) = x'_{ijk}\beta + \tau_i + \eta_j + \zeta_{ij}$. The probability density function for a Gamma (a, b) distribution is $f(x) = a^b x^{a-1}\exp(-bx)/\Gamma(a)$. The α is a shape parameter in the Gamma distribution, which could be assumed common for the entire population (or varied across size levels or AEAs). The distributions on τ_i, η_j, and ζ_{ij} are the same as in the previous model. The hyperparameters are α, β, σ^2_τ, σ^2_η, and σ^2_ζ.

Under both models, the sample design is considered as ignorable because it is an inherent part of the models. That is, the design variables AEA, district size, and school are included in the models. Although these models are specific to the school survey example, the proposed methodology could as easily apply to other hierarchical models fit to data collected on other topics using different sample designs.

3. Hierarchical Bayes Analysis

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

3.1 Prior distributions

In a hierarchical Bayesian framework, we assume mutually independent diffuse prior distributions for the hyperparameters. Let β have a (locally) uniform distribution with $p\left(\beta\right) \propto 1$. Independently $\sigma^2_\tau \sim \text{IG}\left(a_\tau, b_\tau\right)$, $\sigma^2_\eta \sim \text{IG}\left(a_\eta, b_\eta\right)$, and $\sigma^2_\zeta \sim \text{IG}\left(a_\zeta, b_\zeta\right)$, where IG denotes an Inverse-Gamma distribution and a_τ, b_τ, a_η, b_η, a_ζ, and b_ζ are known positive constants. In the Poisson-Lognormal model, it is assumed that $\sigma^2_v \sim \text{IG}\left(a_v, b_v\right)$, where a_v, and b_v are also known positive constants. The constants usually are set to be very small to reflect vague knowledge about the parameters. If a Poisson-Gamma model is employed, the scale parameter α can be assumed to have an independent prior distribution as $\alpha \sim \text{Exponential}\ (1)$. By using the proposed prior distributions, the corresponding posterior (conditional and marginal) distributions are proper.

3.2 Posterior estimates

The posterior distribution of unknown quantities can be approximated by replicative simulates generated using a MCMC algorithm, which was executed using WinBUGS and R in our application. Simulations such as those implemented for this model are commonly implemented for hierarchical models of various specifications in other applications. For each model, $L = 3$ parallel Markov chains were produced. Performance of the MCMC sampling procedure was tested on up to 10 chains, but little difference in results was noted. The convergence of the draws of parameters to their posterior distribution was diagnosed using the Brooks-Gelman-Rubin (BGR) statistic (Gelman et al. 1995). After the convergence had been achieved for all parameters, a subsequence of $R = 1,000$ iterates from each chain was retained for posterior estimation.

Estimates of the posterior mean, variance, and covariance of λ terms are given below. These are followed by the hierarchical Bayesian estimates of μ_{ij}, the average response within stratum (i, j).

The posterior mean and variance of λ_{ijk} under the Poisson-Lognormal model are given by $E(\lambda_{ijk}|y_s) = E\{\exp(x'_{ijk}\beta + \tau_i + \eta_j + \zeta_{ij} + \frac{1}{2}\sigma_v^2)|y_s\}$ and $V(\lambda_{ijk}|y_s) = E\{\exp[2(x'_{ijk}\beta + \tau_i + \eta_j + \zeta_{ij} + \sigma_v^2)]|y_s\} - E^2(\lambda_{ijk}|y_s)$, respectively. These can be estimated using the iterated simulates from MCMC as follows: $\hat{E}\left(\lambda_{ijk}|y_s\right) = \frac{1}{LR}\sum_{l=1}^{L}\sum_{r=1}^{R}[\exp\{x'_{ijk}\beta^{(lr)} + \tau_i^{(lr)} + \eta_j^{(lr)} + \zeta_{ij}^{(lr)} + \frac{1}{2}{\sigma_v^{(lr)}}^2\}]$ and $\hat{V}\left(\lambda_{ijk}|y_s\right) = \frac{1}{LR}\sum_{l=1}^{L}\sum_{r=1}^{R}[\exp\{2(x'_{ijk}\beta^{(lr)} + \tau_i^{(lr)} + \eta_j^{(lr)} + \zeta_{ij}^{(lr)} + {\sigma_v^{(lr)}}^2)\}] - [\hat{E}(\lambda_{ijk}|y_s)]^2$. In the equations above and ones that follow, the superscript (lr) denotes the r^{th} iteration in the l^{th} chain in the retained subsequences. The posterior covariance of λ_{ijk} and $\lambda_{i'j'k'}$ is $C\left(\lambda_{ijk}, \lambda_{i'j'k'}|y_s\right) = C\{\exp(x'_{ijk}\beta + \tau_i + \eta_j + \zeta_{ij} + \frac{1}{2}\sigma_v^2), \exp(x'_{i'j'k'}\beta + \tau_{i'} + \eta_{j'} + \zeta_{i'j'} + \frac{1}{2}\sigma_v^2)|y_s\}$. It can be estimated by $\hat{C}\left(\lambda_{ijk}, \lambda_{i'j'k'}|y_s\right) = \frac{1}{LR}\sum_{l=1}^{L}\sum_{r=1}^{R}\exp\{\left(x_{ijk} + x_{i'j'k'}\right)'\beta^{(lr)} + \tau_i^{(lr)} + \tau_{i'}^{(lr)} + \eta_j^{(lr)} + \eta_{j'}^{(lr)} + \zeta_{ij}^{(lr)} + \zeta_{i'j'}^{(lr)} + {\sigma_v^{(lr)}}^2\} - \hat{E}\left(\lambda_{ijk}|y_s\right)\hat{E}\left(\lambda_{i'j'k'}|y_s\right)$.

If using the Poisson-Gamma model, the posterior mean and variance of λ_{ijk} are $E(\lambda_{ijk}|y_s) = E\{\exp(x'_{ijk}\beta + \tau_i + \eta_j + \zeta_{ij})|y_s\}$ and $V(\lambda_{ijk}|y_s) = E\{\exp[2(x'_{ijk}\beta + \tau_i + \eta_j + \zeta_{ij})(1 + 1/\alpha)]|y_s\} - E^2(\lambda_{ijk}|y_s)$, respectively. They can be estimated using the iterated simulates from MCMC as follows: $\hat{E}\left(\lambda_{ijk}|y_s\right) = \frac{1}{LR}\sum_{l=1}^{L}\sum_{r=1}^{R}[\exp\{x'_{ijk}\beta^{(lr)} + \tau_i^{(lr)} + \eta_j^{(lr)} + \zeta_{ij}^{(lr)}\}]$ and $\hat{V}\left(\lambda_{ijk}|y_s\right) = \frac{1}{LR}\sum_{l=1}^{L}\sum_{r=1}^{R}[\exp\{2(x'_{ijk}\beta^{(lr)} + \tau_i^{(lr)} + \eta_j^{(lr)} + \zeta_{ij}^{(lr)})(1 + 1/\alpha^{(lr)})\}] - [\hat{E}(\lambda_{ijk}|y_s)]^2$. The posterior covariance of λ_{ijk} and $\lambda_{i'j'k'}$ is $C\left(\lambda_{ijk}, \lambda_{i'j'k'}|y_s\right) = C\{\exp(x'_{ijk}\beta + \tau_i + \eta_j + \zeta_{ij}, \exp(x'_{i'j'k'}\beta + \tau_{i'} + \eta_{j'} + \zeta_{i'j'})|y_s\}$. It can be estimated by $\hat{C}\left(\lambda_{ijk}, \lambda_{i'j'k'}|y_s\right) = \frac{1}{LR}\sum_{l=1}^{L}\sum_{r=1}^{R}\exp\{\left(x_{ijk} + x_{i'j'k'}\right)'\beta^{(lr)} + \tau_i^{(lr)} + \tau_{i'}^{(lr)} + \eta_j^{(lr)} + \eta_{j'}^{(lr)} + \zeta_{ij}^{(lr)} + \zeta_{i'j'}^{(lr)}\} - \hat{E}\left(\lambda_{ijk}|y_s\right)\hat{E}\left(\lambda_{i'j'k'}|y_s\right)$.

Let μ_{ij} denote the average number of EP courses taken by twelfth grade students in stratum (i, j) over eight semesters of high school. These quantities are of primary interest in the application. Let s_{ij} and U_{ij} be sets that denote the sample and the population of schools, respectively, in stratum (i, j). Let s_{ijk} and U_{ijk} denote the sample and the population of students in school (i, j, k). The number of students in the stratum is $N_{ij} = \sum_{k\in U_{ij}} N_{ijk}$, where N_{ijk} is the number of students in a school. The average μ_{ij} can be considered as the sum of three terms: $\mu_{ij} = N_{ij}^{-1}\{\sum_{k\in s_{ij}}\sum_{l\in s_{ijk}}\tilde{Y}_{ijkl} + \sum_{k\in s_{ij}}\sum_{l\notin s_{ijk}}\tilde{Y}_{ijkl} + \sum_{k\notin s_{ij}}\sum_{l\in U_{ijk}}\tilde{Y}_{ijkl}\}$, where $\tilde{Y}_{ijkl}|\lambda_{ijk} \sim$ Poisson $(8\lambda_{ijk})$. The first term consists of values observed in the sample adjusted to represent eight semesters. The second term consists of unobserved student values in the selected schools. The third term consists of values from schools not in the sample.

A Bayesian estimator of μ_{ij} is $E\left(\mu_{ij}|y_s\right) = N_{ij}^{-1}\{\sum_{k\in s_{ij}} 8\sum_{l\in s_{ijk}} y_{ijkl}/\omega_{ijkl} + \sum_{k\in s_{ij}} 8(N_{ijk} - n_{ijk})E\left(\lambda_{ijk}|y_s\right) + \sum_{k\notin s_{ij}} 8N_{ijk}E\left(\lambda_{ijk}|y_s\right)\} \equiv N_{ij}^{-1}\{\sum_{k\in s_{ij}} 8\sum_{l\in s_{ijk}} y_{ijkl}/\omega_{ijkl} + l'_{ij}E\left(\lambda|y_s\right)\}$. In the above, $\lambda = \{\lambda_{ijk}\}$ is a parameter vector of Poisson distribution rates for schools in the entire population and $l'_{ij} = \{0, \cdots, 0, \tilde{l}_{ij}, 0, \cdots, 0\}$ is the vector of coefficients for stratum (i, j). In the latter expression, $\tilde{l}_{ij} = \{l_{ijk}\}_{k\in U_{ij}}$ is the vector of values l_{ijk} in stratum (i, j). The value l_{ijk} equals $8(N_{ijk} - n_{ijk})$ if $k \in s_{ij}$. It equals $8N_{ijk}$ if $k \notin s_{ij}$. The proposed HB estimator of μ_{ij} is $\hat{\mu}_{ij} = N_{ij}^{-1}\{\sum_{k\in s_{ij}} 8\sum_{l\in s_{ijk}} y_{ijkl}/\omega_{ijkl} + l'_{ij}\hat{E}\left(\lambda|y_s\right)\}$. The posterior variance of $\hat{\mu}_{ij}$ is $V\left(\mu_{ij}|y_s\right) = N_{ij}^{-2}\{l'_{ij}V\left(\lambda|y_s\right)l_{ij}\}$, which can be estimated by plugging $\hat{V}\left(\lambda|y_s\right)$.

4. Benchmarked HB Model Selection

Model selection 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. The posterior predictive methods such as posterior predictive p-value, L-criterion, and deviance information criterion (DIC) are commonly used for model selection.

The posterior predictive p-value (Meng 1994, Gelman, Meng, and Stern 1996) measures the tail-area probability in the posterior predictive distribution based on a discrepancy measure which could be a test statistic or more generally involving the unknown nuisance parameters from the model. A small p-value provides "evidence" against the assumed model. The L-criterion (Laud and Ibrahim 1995) measures the performance of a model by evaluating expected posterior predictive errors and also calibrating on the uncertainty associated with the criterion value. Hoeting and Ibrahim (1998) defined a calibration comparison score (CCS) 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 CCS is preferred. The DIC (Spiegelhalter et al. 2002) calculates the sum of the posterior mean deviance and the effective number of parameters, which is a Bayesian measure of fit or adequacy penalized by model complexity p_D. A model with smallest DIC value is tend to be selected. All the methods allow the use of non-informative prior distributions, which could be desirable for model checking and comparison especially at the preliminary stage of model exploration. They are easy to implement in many hierarchical models for which MCMC simulation can be performed. The posterior predictive p-value method can use multiple discrepancy measures to evaluate a model based on just one set of simulation. On the other hand, a shortcoming of all the methods is that they can be very conservative and have low power due to the double use of data. This shortcoming has been addressed and discussed by Draper (1996) and Bayarri and Castellanos (2007). Alternative methods of avoiding the double use of data include the partial posterior predictive p-value (Bayarri and Castellanos 2007) and cross-validated posterior predictive checking (Stern and Cressie 2000; Larsen and Lu 2007). These alternatives are not considered here but could be studied in the future.

In studies in which small areas are of interest but have small sample sizes, it can happen that there is a big enough sample for producing reliable estimates for larger regions composed of groups of small areas. You, Rao, and Dick (2004) proposed "benchmarking" the HB estimators of small areas so that the benchmarked HB estimators will add up to the direct estimators in larger regions. For example, in the ISBE EP survey, we can benchmark the HB estimators for AEAs (strata) in a certain size level so that the sum of the benchmarked HB estimators over all strata in the size level equals the direct estimator of the size level. The benchmark property with respect to the size level direct estimator is given by $\sum_j N_{ij}\hat{\mu}_{ij}^{BHB} = \sum_j N_{ij}\hat{\bar{y}}_{ij}$, where $i \in \{\text{size level: } 1 = \text{large}; 2 = \text{medium}; 3 = \text{small}\}$, $j \in \{12 \text{ AEAs}\}$, and $\hat{\bar{y}}_{ij}$ denotes the direct estimate of the population mean for stratum (i, j). In particular, the raking-benchmarked HB (RBHB) estimator for stratum (i, j) is $\hat{\mu}_{ij}^{RBHB} = \hat{\mu}_{ij}^{HB}\frac{\sum_j N_{ij}\hat{\bar{y}}_{ij}}{\sum_j N_{ij}\hat{\mu}_{ij}^{HB}}$.

The variation associated with the BHB estimator under the assumed model can be measured by the posterior mean square error (PMSE): $PMSE(\hat{\mu}_{ij}^{BHB}) = E[(\hat{\mu}_i^{BHB} - \mu_i)^2|y^{obs}]$. When the BHB estimators are the same as the HB estimators, $E[(\hat{\mu}_i^{HB} - \mu)|y^{obs}] = 0$ under the assumed model and the PMSE is equal to the posterior variance. PMSE of the BHB estimator can be estimated as $PMSE\left(\hat{\mu}_{ij}^{BHB}\right) = V\left(\mu_{ij}|y^{obs}\right) + \left(\hat{\mu}_{ij}^{BHB} - \hat{\mu}_{ij}^{HB}\right)^2$, which is the sum of posterior variance $V\left(\mu_{ij}|y^{obs}\right)$ and a bias correction term $\left(\hat{\mu}_{ij}^{BHB} - \hat{\mu}_{ij}^{HB}\right)^2$.

Benchmarked HB estimators are design consistent in larger regions, which is an attractive property. Due to benchmarking the BHB estimators should be more robust to model failure than the HB estimators. When the model is misspecified, benchmarking could correct the bias of the HB estimator to some degree. The PMSE derived under the model, however, would be inflated correspondingly due to the bias correction. The farther the specified model is from the true model, the more serious the inflation of PMSE there could be. Therefore, a big inflation of PMSE can suggest possible model inadequacy. We can measure the degree of discrepancy between the model and the observed data based on the degree to which benchmarking inflates the PMSE of the HB estimator. Below we describe using benchmarked HB estimation results and PMSE inflation for the purpose of model selection.

Let $\Delta = (\text{PMSE}^{BHB} - \text{PMSE}^{HB})/\text{PMSE}^{HB}$, which is the relative change of PMSE due to benchmarking. Equivalently $\Delta = (\hat{\mu}^{BHB} - \hat{\mu}^{HB})^2/E[(\hat{\mu}^{HB} - \mu)^2|y^{obs}]$. There are at least a couple of ways to translate this measure into a discrepancy measure. Let h index small areas (strata) and Δ_h be the value of Δ for area h. We define $D_{1;BMB}$ as the proportion of small areas having Δ_h bigger than a certain value, say δ. The discrepancy can be expressed as $D_{1;BHB}(y, \theta) = H^{-1}\sum_{h=1}^{H} I_{\Delta_h > \delta}$. For example, if we choose $\delta = z_{0.975}^2$ where $z_{0.975} = \Phi^{-1}(0.975)$ is the 97.5% percentile of a standard normal distribution, then $\Delta_h > \delta$ is equivalent to $|\hat{\mu}_h^{BHB} - \hat{\mu}_h^{HB}| > z_{0.975}\sqrt{V(\mu_h|y^{obs})}$, which means we are measuring the number of strata having benchmarked HB estimates falling out of the 95% asymptotic normal posterior predictive intervals of the HB estimators. Alternatively, we can quantify the overall inflation of PMSE for BHB versus HB as $D_{2;BHB}(y, \theta) = H^{-1}\sum_{h=1}^{H}\Delta_h$, which is the average relative change of PMSE over all small areas.

For a given discrepancy $D(y, \theta)$, the posterior predictive check will be based on the comparison between the predictive discrepancy $D(y^{pred}, \theta)$ and the realized discrepancy $D(y^{obs}, \theta)$. The posterior predictive p-values based on the discrete discrepancy $D_{1;BHB}(y, \theta)$ and on the continuous discrepancy $D_{2;BHB}(y, \theta)$ measure are defined as $p_{post,1;BHB} = Pr(D_{1;BHB}(y^{pred}, \theta) \geq D_{1;BHB}(y^{obs}, \theta)|y^{obs})$ and $p_{post,2;BHB} = Pr(D_{2;BHB}(y^{pred}, \theta) > D_{2;BHB}(y^{obs}, \theta)|y^{obs})$. Estimating these posterior predictive discrepancies can be accomplished from MCMC output with a little effort. In particular, for each value of θ and y^{pred}, one must compute $D_{1;BHB}$ and $D_{2;BHB}$. Thus, both the HB and BHB estimates themselves are computed using MCMC for each y^{pred}.

5. Simulation

To illustrate the performance of the proposed estimators and model comparison methods, we simulated a finite population of EP courses taken by twelfth grade students from Iowa's public high schools from a Poisson loglinear model with random effects from size levels and AEAs: $y_{ijkl}|\lambda_{ijk} \sim \text{Poisson }(\lambda_{ijk})$, $\log(\lambda_{ijk}) = \beta_0 + \beta_1 x_{ijk;1} + \tau_i + \eta_j$. Students in the simulated population were assumed to attend the same number of semesters so that the exposure variable of the attendance of semesters was excluded. The enrollment size in twelfth grade was used as an auxiliary variable x_1 in generating the population data set. 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 Poisson-loglinear models consisting of different combinations of auxiliary variables and random effects (Lu and Larsen 2007a) were fit to the sample data. Among these models, model 3 is the model from which the population was simulated.

Table 1 shows the results of comparing the seven models using the three existing methods discussed in Section 4 and the method of posterior predictive checking based on the newly developed benchmark discrepancies. The posterior predictive p-value using the χ^2-discrepancy $\sum_{i,j,k,l\in s}(y_{ijkl} - \lambda_{ijk})^2/\lambda_{ijk}$ is denoted $p_{post;\chi^2}$. The p-values based on the benchmark discrepancies are denoted $p_{post,1;BHB}$ and $p_{post,2;BHB}$, respectively. According to $P_{post;\chi^2}$, models 1, 2, 5 and 6 show strong evidence of model failure. Models 3, 4 and 7 have no indication of model inadequacy based on the same measure. Of these, model 3 is the most parsimonious. When using the L-criterion, model 7 has the smallest criterion value. Calibrated by the standard deviation of the criterion value under model 7, the CCSs for models 1, 2, and 6 are larger than the value of 3, which is too extreme. Model 5 is the smallest model with a not extreme CCS. Among models having small DIC value, model 3 has the smallest number of effective parameters. When using benchmarked HB model selection based on discrepancy $D_{1;BHB}(y,\theta)$, only models 2 and 6 end up with extreme $p_{post,1;BHB}$ values. The other models show no significant incompatibility between model and data. Of these, model 1 is the winner according to the parsimonious rule. When using the discrepancy $D_{2;BHB}(y,\theta)$, models 1, 2 and 6 have very extreme $p_{post,2;BHB}$ values. The other models have shown no extreme patterns of the observed data relative the replicate predictive data in terms of the discrepancy $D_{2;BHB}(y,\theta)$. Model 5 is the simplest model among models with $p_{post,2;BHB}$ bigger than 0.05. The $p_{post;\chi^2}$ and DIC criteria successfully choose the true model. The L-criterion and $p_{post,2;BHB}$ select 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 slightly larger absolute relative bias (ARB) to the realized finite population mean and higher posterior mean square error (PMSE) in most of strata, but the estimates under two models are still very close. The ARB is defined as the absolute value of the relative bias of the estimate over the realized finite population value. The $p_{post,1;BHB}$ fails to detect the significant model inadequacy of model 1. The discrepancy $D_{1;BHB}(y,\theta)$, being discrete, probably loses some power relative to quantitative $D_{2;BHB}(y,\theta)$. Basically, all the Bayesian model comparison methods discussed above except discrepancy $D_{1;BHB}(y,\theta)$ work well in selecting an appropriate model for further analysis. One lesson of this work is that referring to multiple criteria if practically feasible should be helpful in making a good decision. See Larsen and Lu (2007) for another example in this spirit.

Table 1: Model selection results for the simulation. Model 3 is the true model. $p_{post:\chi^2}$=posterior predictive p-value based on the χ^2 discrepancy. *CCS*=calibration comparison score for the L-criterion value. *DIC*=deviance information criterion. p_D=effective number of parameters. $p_{post,1;BHB}$=posterior predictive p-value based on the discrepancy $D_{1;BHB}(y,\theta)$. $p_{post,2;BHB}$=posterior predictive p-value based on the discrepancy $D_{2;BHB}(y,\theta)$. Bold values indicate models that cannot be declared inappropriate.

	$p_{post;\chi^2}$	*CCS*	*DIC*	p_D	$p_{post,1;BHB}$	$p_{post,2;BHB}$
M1	0.000	5.51	20410	2.22	**0.455**	0.000
M2	0.000	3.92	20160	3.85	0.035	0.000
M3	**0.125**	**0.03**	**19500**	14.93	**1.000**	**1.000**
M4	**0.146**	**0.01**	**19500**	20.70	**1.000**	**1.000**
M5	0.011	**0.67**	19610	13.78	**1.000**	**1.000**
M6	0.000	5.40	20350	4.05	0.019	0.000
M7	**0.146**	**0.00**	**19500**	24.98	**0.978**	**0.997**

In a preliminary study (Lu and Larsen 2006), we compared direct survey estimates based on a ratio estimator and on the Horvitz-Thompson estimator for estimates within strata. The ratio estimator produced estimates with smaller variance and mean square error (MSE) in the Monte Carlo study. Then we compare the model-based HB estimator with the design-based ratio estimator based on the absolute relative bias (ARB) and root mean square error (RMSE) for individual strata (Lu and Larsen 2007a). The MSE of the ratio estimator is estimated through Monte Carlo simulation. For the single randomly selected sample in the simulation, the ratio estimator produces consistently larger ARBs for most of strata and shows much higher variation and produces consistently larger RMSE than 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 between the other two.

Also 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. In the case of our application, the collapsing strata synthetic variance (CSSV) estimator significantly overestimated the variances in small areas. The collapsed strata restricted generalized variance function synthetic variance (CRGVFSV) estimation method (Lu and Larsen 2007b, 2008) 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. The BHB estimator has larger PMSE than HB estimator due to the benchmarking procedure, but it still produces smaller variance than the ratio estimator for most of the strata, especially for strata with a single PSU in the sample.

6. ISBE Survey Data

The actual survey data were collected in 2005 from 51 sample schools in 11 AEAs. Hierarchical Bayesian (HB) estimation are applied to analyze the survey data. We fit both Poisson-Lognormal and Poisson-Gamma models to the survey data and compare models with different combinations of auxiliary variable and random effects. Neither of the L-criterion and DIC methods can tell even a slight difference between the models. The posterior predictive p-values based on nine discrepancy measures are calculated under each model structure. The p-values based on all measures do not show any difference between the models under the same model structure. For example, the nine discrepancies used for Poisson-Gamma model are: $D_1 = \sum_i \sum_j \sum_k \bar{y}_{ijk} = \sum_i \sum_j \sum_k \left(\sum_{l \in s_{ijk}} y_{ijkl} / \sum_{l \in s_{ijk}} \omega_{ijkl} \right)$, $D_2 = \sum_i \sum_j \sum_k \sum_{l \in s_{ijk}} \frac{(y_{ijkl} - \omega_{ijkl}\lambda_{ijk})^2}{\omega_{ijkl}\lambda_{ijk}}$, $D_3 = \sum_i \sum_j \sum_k \frac{(\bar{y}_{ijk} - \gamma_{ijk})^2}{\gamma_{ijk}/\omega_{ijk} + \gamma_{ijk}^2/\alpha}$, $D_4 = \max_{i,j,k,l \in s} (y_{ijkl}/\omega_{ijkl})$, $D_5 = \min_{i,j,k,l \in s} (y_{ijkl}/\omega_{ijkl})$, $D_6 = \max_{i,j,k,l \in s} (y_{ijkl}/\omega_{ijkl}) - \min_{i,j,k,l \in s} (y_{ijkl}/\omega_{ijkl})$, $D_7 = \max_{i,j,k \in s} \bar{y}_{ijk}$, $D_8 = \min_{i,j,k \in s} \bar{y}_{ijk}$, and $D_9 = \max_{i,j,k \in s} \bar{y}_{ijk} - \min_{i,j,k \in s} \bar{y}_{ijk}$, where $\omega_{ijk} = \sum_{l \in s_{ijk}} \omega_{ijkl}$. Measures D_2 (the first level χ^2-discrepancy) and D_5 (the minimum) indicate an inadequacy of the Poisson model at the school level. Accordingly, we would like to consider a more complex model structure such as a zero-inflated Poisson model or a mixture Poisson model for capturing the distribution pattern of students within schools in the future. Results for the Poisson-Lognormal model structure are the same. The one difference is that the p-values based on measure D_3 (the second level χ^2-discrepancy) under the Poisson-Lognormal models are consistently larger than those under Poisson-Gamma models. This difference suggests distinct performance of two model structures.

Consider, for example, the models with only AEA effects. The Poisson-Lognormal model produces significantly higher second level χ^2-discrepancy measure (D_3) in the posterior predictions than it does for the realized data. Using the same explanatory variables, the Poisson-Gamma model produces posterior predictive discrepancies fairly evenly around the realized discrepancy value. This indicates the data generated from the fitted Poisson-Lognormal model have more variation than the observed data. The data generated from the fitted Poisson-Gamma model show no extreme pattern compared with the acutal data. The underlying reason for this is the higher skewness for the Poisson-Gamma distribution.

Figure 1 shows the probability density functions of Gamma $(\frac{1}{e-1}, \frac{1}{(e-1)e^{1/2}})$, Lognormal$(0, 1)$ and Lognormal$(0, 2.5)$ distributions. The Gamma $(\frac{1}{e-1}, \frac{1}{(e-1)e^{1/2}})$ and Lognormal$(0, 1)$ have exactly the same mean $(e^{1/2})$ and variance $(e(e-1))$, but the shapes of the two distributions are different. The Gamma distribution is more skewed towards zero with much bigger probability for values around zero. The Lognormal$(0, 2.5)$ with larger variance than the other lognormal distribution matches the Gamma distribution except around the area very close to zero. Therefore, if the data were actually generated from a Poisson-Gamma distribution, to fit the data using Poisson-Lognormal model would result in an estimated model with larger variance.

Table 2 shows the posterior predictive p-values based on the discrepancy $D_{2;BHB}(y, \theta)$. The model with only an auxiliary variable and no random effects has the biggest average inflation of PMSE under either model structure. The models including only size effect with and without the auxiliary variable also have much larger realized discrepancy values than other models that include an AEA effect. To see whether the realized discrepancy is extreme or not under the assumed model, we compare the $p_{post,2;BHB}$ values. For the Poisson-Gamma models, the $p_{post,2;BHB}$ values show that there are relatively smaller chances to have more extreme discrepancy values in the replicated predictions than for the actual data under the three models with significantly larger realized discrepancies. The relatively small $p_{post,2;BHB}$ values indicate these three models are less compatible with the actual data than other Poisson-Gamma models in terms of producing model-based mean estimates that differ from the reliable direct estimates.

None of the Poisson-Lognormal models has an extreme $p_{post,2;BHB}$ value. Even the three models (with *Aux.*, *Aux. & size*, and *Size*) that have quite large realized discrepancy measures (greater than 3) have $p_{post,2;BHB}$ values no less than 0.25. So the posterior predictive check based on the p-value for the given discrepancy measure shows no evidence of model incompatibilities with the actual data for all Poisson-Lognormal models. Since for models which have the same basic model structure and all fit the data well we usually choose the most parsimonious model for further analysis. By only referring to the $p_{post,2;BHB}$ values, we would like to select the Poisson-Gamma model with only an AEA effect or a Poisson-Lognormal model with only an auxiliary variable. Further, to compare these two models with different model structures, we computed $(D_{2;BHB}(y^{pred}, \theta) - D_{2;BHB}(y^{obs}, \theta))$ for the two models. The posterior predictive discrepancies under the Poisson-Lognormal model have a much wider spread than those under the Poisson-Gamma model. This indicates that there is generally more variation in the fitted Poisson-Lognormal model than the Poisson-Gamma model, which is consistent with our previous finding based on examining the second level χ^2-discrepancy in the posterior predictions. This also explains why the Poisson-Lognormal models with very large realized discrepancy values do not have extreme $p_{post,2;BHB}$ values. Therefore, we prefer the Poisson-Gamma model with only an AEA effect, which fits the data well with a relatively simple model structure and produces reliable posterior estimates of means for small areas.

The analysis of the Iowa survey data gives an example in which the L-criterion and the DIC method have less ability to choose models than the proposed method. The posterior predictive p-values using different discrepancy measures show advantages of detecting incompatibilities between the data and different parts of the model. This also was seen in Larsen and Lu (2007). The posterior predictive check based on the newly developed discrepancy measure of the inflation of PMSE due to benchmarking the HB estimator did much better in assessing overall fit of models than other examined discrepancies.

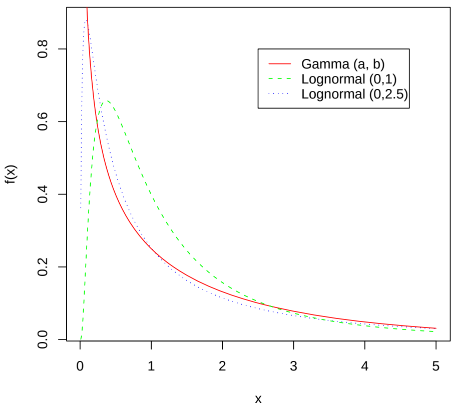

Figure 1: The probability density function of Gamma ($a = \frac{1}{e-1}, b = \frac{1}{(e-1)e^{1/2}}$), Lognormal$(0, 1)$ and Lognormal$(0, 2.5)$ distributions.

Table 2: The posterior predictive p-value $p_{post,2;BHB}$ based on the discrepancy $D_{2;BHB}(y, \theta)$ and the realized discrepancy $D_{2;BHB}(y^{obs}, \theta)$ using the Poisson-Gamma and Poisson-Lognormal models for the Iowa survey data. Bold realized discrepancy values indicate large deviation of HB from direct estimate in large regions, which suggest potential model inadequacy. Bold p values indicate relatively small probability of observing more extreme predictive data in terms of the discrepancy measure than the observed data, which suggest more incompatibilities between the data and the models.

Models	*Poisson-Gamma*		*Poisson-Lognormal*	
	$p_{post,2;BHB}$	$D_{2;BHB}(y^{obs}, \theta)$	$p_{post,2:BHB}$	$D_{2;BHB}(y^{obs}, \theta)$
Aux.	**0.063**	**6.783**	0.265	**5.990**
Aux. & Size	**0.115**	**4.065**	0.649	**3.850**
Aux. & AEA	0.345	0.644	0.879	0.733
Aux. & Size & AEA	0.389	0.362	0.639	0.474
Aux. & Size & AEA & Inter.	0.557	0.293	0.994	0.099
Size	**0.057**	**3.407**	0.315	**3.255**
AEA	0.475	0.588	0.762	0.987
Size & AEA	0.381	0.288	0.652	0.333
Size & AEA & Inter.	0.468	0.250	0.952	0.152

7. Conclusion

In studies involving small areas with very small sample sizes, using a model-based estimator to produce reliable estimates of small area quantities is desirable. A survey on transcripts of Iowa's public high school students motivated an examination of small area estimation through model-based inference. A hierarchical Bayes (HB) estimator was proposed to obtain the estimates of the average number of EP courses taken by twelfth grade high school students for strata defined by district size and AEA and for populations of aggregations of strata. When an appropriate model is used, the HB estimator is shown to outperform the ratio estimator in a simulation study by borrowing strength across strata and making better use of auxiliary information in terms of producing consistently smaller absolute relative bias (ARB) (relative to the realized finite population mean) and root mean square error (RMSE) for individual strata.

Effective model selection is crucial in the HB analysis. The issue of model selection not only includes selecting proper model structure but also selecting covariate variables and proper forms of transformations of the variables. A HB posterior predictive model comparison method utilizing benchmarking is developed and shown to have the power to choose appropriate models in both a simulation study and a real data analysis. The proposed method was compared to Bayesian model comparison based on the posterior predictive p-values using multiple discrepancy measures, the L-criterion, and the deviance information criterion. The last two methods did a reasonable job in the simulation study but showed less ability to detect inadequate models in analyzing the real data. The posterior predictive p-value using multiple discrepancy measures showed advantages in comparing models which are undistinguishable using the other two methods. The proposed discrepancy which measures the inflation of PMSE due to benchmarking the HB estimator outperforms other examined discrepancy measures in terms of evaluating the overall fit of models.

Future study will examine methods of choosing 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 studies, since it is practically inefficient or impossible to compare all possible models with various combinations of variables, we also hope to explore methods to narrow the range of candidate models. Further, given the similarity of overall performance of many models, Bayesian model averaging in the small area context might be another option for future study.

Acknowledgments

The authors wish to thank Iowa's State Board of Education and the American Education Research Association for partial support of this work. The opinions expressed in this article are those of the authors and do not necessarily belong to anyone else.

REFERENCES

Bayarri, M.J., and Castellanos, M.E. (2007), ”Bayesian checking of the second levels of hierarchical models (with discussion)”, *Statistical Science*, **22**, No. 3, 322-343.

Draper D. (1996), ”Comment: On posterior predictive p-values, discussion of Gelman, A., Meng, X.L., and Stern, H. Posterior predictive assessment of model fitness via realized discrepancies”, *Statistica Sinica*, **6**, 760-767.

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.

Larsen, M.D., and Lu, L. (2007), ”Comment: Bayesian Checking of the Second Level of Hierarchical Models: Cross-Validated Posterior Predictive Checks Using Discrepancy Measures”, *Statistical Science*, **22**, No. 3, 359-362.

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.

Lu, L., and Larsen, M.D. (2007a), ”Small Area Estimation in a Survey of High School Students in Iowa”, *Proceedings of the Survey Research Methods Section*, American Statistical Association.

Lu, L., and Larsen, M.D. (2007b), ”Variance Estimation in a High School Student Survey with One-Per-Stratum Strata”, *Proceedings of the Third International Conference on Establishment Surveys (ICES-III).*

Lu, L., and Larsen, M.D. (2008), ”Variance estimation for One-Per-Stratum Strata”, *Journal of Official Statistics*, Revise and resubmit.

Meng, X.L. (1994), ”Posterior predictive p-values”, *Annals of statistics*, **22**, 1142-1160.

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.

Stern, H.S., and Cressie, N. (2000), ”Posterior predictive model checks for disease mapping models”, *Statistics in Medicine*, **19**, 2377-2397.

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.