Statistical methods in a high school transcript survey

by

Lu Lu

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Statistics

Program of Study Committee:
Michael D. Larsen, Major Professor
Jean Opsomer
Taps Maiti
Fred Lorenz
Amy Froelich

Iowa State University

Ames, Iowa

2009

TABLE OF CONTENTS

ABSTRACT xii

CHAPTER 1. Introduction 1

CHAPTER 2. Transcript Survey Study by Iowa's State Board of Education 5

2.1 Sample Design 6

2.2 Design and Estimation Options in the Survey 10

2.2.1 Direct Estimators 10

2.2.2 An Invariant And A Non-invariant Design 13

2.2.3 A Cost Evaluation 14

2.3 Summary 16

CHAPTER 3. Variance Estimation in a One-per-stratum Design 18

3.1 Existing approaches for variance estimation in One-Per-Stratum Designs 19

3.2 Adapted Methods for Variance Estimation in One-Per-Stratum Strata 20

3.2.1 Collapsing strata synthetic estimation of stratum variances 20

3.2.2 Modeling and generalized variance functions 21

3.3 Simulation Studies 23

3.3.1 Simulation to compare direct variance estimation and GVF adjustment in general 24

3.3.2 Simulation to study CSSV and GVF adjustment in single PSU designs . 28

3.4 Results for the Iowa Employment Preparation Survey 31

3.5 Summary and Discussion 35

CHAPTER 4. Small Area Estimation using Hierarchical Bayesian Analysis 36

4.1 Small Area Estimation and Existing Methods 36

4.2 Generalized Linear Mixed Models 37

4.3 Hierarchical Bayes Analysis 39

4.3.1 Prior distributions 39

4.3.2 MCMC sampling 39

4.3.3 Posterior estimates 41

4.4 Illustration 44

4.5 ISBE Survey Data Analysis 48

4.6 Summary and Discussion 55

CHAPTER 5. Hierarchical Bayesian Model Selection Using Benchmarking 63

5.1 Existing methods for Bayesian model checking and model comparison 64

5.1.1 Posterior Predictive P-Value 65

5.1.2 L-Criterion 70

5.1.3 Deviance Information Criterion 72

5.2 Benchmarked HB Model Comparison 75

5.2.1 Benchmarked HB estimator 76

5.2.2 The use of benchmarked discrepancy in posterior predictive model selection 77

5.3 Illustration 79

5.4 Model Selection in the Analysis of the ISBE Survey Data 86

5.5 Concluding Remarks 96

CHAPTER 6. Conclusions and Future Study 101

6.1 Conclusions 101

6.2 Future Study 103

APPENDIX Derivations for the Posterior Mean and Variance-Covariance of School Level Poisson Rate λ for Hierarchical Bayesian Models 106

BIBLIOGRAPHY 109

ACKNOWLEDGEMENTS 113

LIST OF TABLES

Table 2.1 The numbers of school districts within each stratum in the sample design and in the population. In each pair of parentheses, the number in front of the semicolon is the number of school districts in the population of the stratum, and the number after the semicolon is the number of districts in the sample. 9

Table 2.2 The standard deviation (SD) and the root of mean squared error (RMSE) of total estimates using HT and ratio estimators for three aggregations. 12

Table 2.3 CVEs (displayed in %) of Horvitz-Thompson and ratio estimators under invariant and non-invariant designs for three aggregations. 14

Table 2.4 Empirical percentiles over 1,000 simulations of the relative decrease in variance estimates due to adding more schools to the sample with fixed total cost per school. Decrease is relative to a sample with 60 schools. Variance estimates for the cases of 60 and 65 sample schools used the collapsed strata estimator. Cost factor per school is the reduction in the total number of students in the sample per additional school. IQR is the inter-quantile range of variance reductions. 15

Table 3.1 The estimates of coefficient of variation (cve's) (measured in %) of total estimates using DIR (direct variance estimation), GVF (traditional model without adjustment), AGVF (GVF with adjustment of coefficients), AGVF(log) (AGVF using log-transformed model), ARGVF (restricted GVF with adjustment of coefficients), and CRGVFSV (collapsing strata restricted GVF with synthetic variance redistribution). Results are based on 1,000 replicated samples using the Horvitz-Thompson (HT) estimator. 25

Table 3.2 The estimated coefficients of variation (cve's) (measured in %) of total estimates using CSSV and three adjusted GVF methods. Results are based on 1,000 replicated samples, estimation using the ratio estimator 31

Table 3.3 Number of confidence intervals obtained by using CSSV and CRGVFSV estimation out of 1,000 samples covering totals for strata with medium size districts. 33

Table 4.1 The numbers of schools districts within each stratum in the actual sample data and in the population. "M" means there are school districts sampled in the stratum which refused to participate in the survey. "None" means no school districts existing in the stratum. 50

Table 4.2 Posterior estimates of parameters in the Poisson-Gamma model with size and AEA random effects and no auxiliary variable: $\log(\gamma_{i,j,k}) = \beta_0 + \tau_i + \eta_j$. The bold figures are the 95% credible interval bounds for the intervals that exclude 0. The $\tau_i, i \in \{\text{large, medium, small}\}$ denote size random effects. The $\eta_j, j = 1, 4, 8, 9, 10, 11, 12, 13, 14, 15, 16, 267$ represent AEA random effects. The subscript index the actual size levels and AEAs. 58

Table 4.3 Posterior estimates of parameters in the Poisson-Gamma model with AEA random effect and an auxiliary variable with random coefficient for size levels: $\log(\gamma_{i,j,k}) = \beta_0 + \beta_{1,i}x_{i,j,k} + \eta_j$. 59

Table 4.4 The HB estimates of stratum means and the coefficient of variation (CV) of the HB estimates for Poisson-Gamma model with size and AEA random effects and no auxiliary variable (Model 1): $\log(\gamma_{i,j,k}) = \beta_0+\tau_i+\eta_j$ and Poisson-Gamma model with a random coefficient for size levels and AEA random effect (Model 2): $\log(\gamma_{i,j,k}) = \beta_0 + \beta_{1,i}x_{i,j,k} + \eta_j$. The CVs for the ratio estimator are calculated from the variance estimates obtained using the collapsed strata synthetic variance (CSSV) and the collapsed strata generalized variance function synthetic variance (CRGVFSV) methods, which are denoted by CV_1 and CV_2 respectively. 60

Table 4.5 Posterior estimates of parameters in Poisson-Lognormal model with size and AEA random effects and no auxiliary variable: $\log(\lambda_{i,j,k}) = \beta_0 + \tau_i + \eta_j + v_{i,j,k}$. 61

Table 4.6 Posterior estimates of stratum means for Poisson-Lognormal model with size and AEA random effects and no auxiliary variable: $\log(\lambda_{i,j,k}) = \beta_0 + \tau_i + \eta_j + v_{i,j,k}$. 62

Table 5.1 Seven models reflecting different assumptions about the between-school variation in log-rate of taking EP courses and the involvement of co-variable variables in the illustrative example. 81

Table 5.2 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 (ϕ_m) for the L_m statistic. *DIC* = deviance information criterion. p_D = effective number of parameters. $p_{post}^{1;BHB}$ = posterior predictive p-value based on the discrepancy $D_1^{BHB}(\boldsymbol{y}, \boldsymbol{\theta})$. $p_{post,2}^{BHB}$ = posterior predictive p-value based on the discrepancy $D_2^{BHB}(\boldsymbol{y}, \boldsymbol{\theta})$. Bold values indicate models that can be declared inappropriate. 83

Table 5.3 Discrepancy measures used in posterior predictive checking for candidate models applied to the Iowa survey data. In the measures, $\omega_{ijk} = \sum_{l \in s_{ijk}} \omega_{ijkl}$. 88

Table 5.4 The posterior predictive p-values for candidate models based on nine discrepancies described in Table 5.3 applied to the Iowa survey data. . 98

Table 5.5 The posterior predictive p-values $p^{BHB}_{post,1}$ and $p^{BHB}_{post,2}$ based on the discrepancies $D^{BHB}_1(\boldsymbol{y}^{obs}, \boldsymbol{\theta})$ and $D^{BHB}_2(\boldsymbol{y}^{obs}, \boldsymbol{\theta})$ and the realized discrepancies $D^{BHB}_1(\boldsymbol{y}^{obs}, \boldsymbol{\theta})$ and $D^{BHB}_2(\boldsymbol{y}^{obs}, \boldsymbol{\theta})$ 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. 99

LIST OF FIGURES

Figure 2.1 The structure of the public educational system in Iowa. 6

Figure 2.2 The geographical division of 12 AEAs in Iowa. 7

Figure 2.3 The sample allocation in the design of 60 sample schools. 8

Figure 3.1 The rate of confidence intervals obtained using direct and CRGVFSV variance estimation covering the true total over 1,000 samples under the designs of $n_1 = 20$ and $n_2 = 5$. 26

Figure 3.2 The standardized root of mean squared error (SRMSE) of variance estimates, defined in (3.6), using direct and CRGVFSV variance estimation based on 1,000 samples under the design of $n_1 = 20$ and $n_2 = 5$. 27

Figure 3.3 The coverage rate of confidence intervals using CSSV and CRGVFSV methods for one-PSU strata over 1,000 samples under a PPS design. Strata 1 to 30 (displayed) each have one PSU selected. Twenty strata (not shown) have two PSUs selected and are used to fit the RGVF. . 29

Figure 3.4 The estimated coefficients of variation (cve's) of variance estimates using CSSV and CRGVFSV methods for one-PSU strata over 1,000 samples under a PPS design. Strata 1 to 30 (displayed) each have one PSU selected. Twenty strata (not displayed) have two PSUs selected and are used to fit the RGVF. 30

Figure 3.5 Average of standard errors of the ratio estimator using CSSV and CRGVFSV variance estimators for strata of medium districts in the case of 60 sample schools. SD is the average of the standard deviations of mean estimates over 1,000 samples. 32

Figure 3.6 Empirical percentiles of the width of confidence intervals obtained by collapsing strata synthetic variance (CSSV) and CRGVF estimation over 1,000 simulations for strata with medium size districts using the ratio estimator. 34

Figure 4.1 The Absolute Relative Bias (ARB) of ratio and HB estimators for strata of medium districts based on a sample data set from a single simulated finite population. 46

Figure 4.2 The root of mean squared error (RMSE) of the ratio estimator and the root of posterior mean squared error (RPMSE) of the HB estimator for strata of medium districts. The RMSE of ratio estimate was obtained based on 1,000 simulated samples. 47

Figure 4.3 The standard errors (SEs) of the ratio estimate and RPMSE of the HB estimate for strata of medium districts based on a sample data set from a simulated finite population. The SEs of ratio estimate was obtained by using collapsed strata synthetic variance (CSSV) and collapsed strata generalized variance function synthetic variance (CRGVFSV) methods, denoted by RA1 and RA2 respectively. 48

Figure 5.1 The absolute relative biases (ARBs) of HB estimates under the true model (model 3) and an inadequate model (model 2) in the illustrative example. 84

Figure 5.2 The absolute relative biases (ARBs) of BHB estimates under the true model (model 3) and an inadequate model (model 2) in the illustrative example. 85

Figure 5.3 The root of posterior mean square error (RPMSE) of the HB and the BHB estimates under the true model (model 3) on the left and under an inadequate model (model 2) on the right in the illustrative example. 86

Figure 5.4 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. 90

Figure 5.5 The scatter plot of sorted ($D_2^{BHB}(\boldsymbol{y}^{pred}, \boldsymbol{\theta}) - D_2^{BHB}(\boldsymbol{y}^{obs}, \boldsymbol{\theta})$) values under Poisson-Gamma model with only AEA effect and no covariate variable and Poisson-Lognormal model with only an auxiliary variable and no random effects involved. 94

Figure 5.6 The overlaid histograms of predictive discrepancy values with vertical lines representing the realized discrepancies for Poisson-Gamma models. In the legends, "PG", "FC", "RC", and "RE" stand for "Poisson-Gamma" model, "fixed coefficient", "random coefficient", and "random effect", respectively. For example, "PG_REAEA" represents the Poisson-Gamma model with only AEA random effect and no auxiliary variable; "PG_RCSize_REAEA" denotes the Poisson-Gamma model with AEA random effect and an auxiliary variable with random coefficient for size levels. 100

ABSTRACT

In complex surveys that involve stratification and clustering structures, given the budget, time and resource restrictions, the surveys are usually designed to produce specific accuracy of direct estimation at high levels of aggregation. Sample sizes for small geographical areas or subpopulations are typically small such that direct estimates in these areas are very unreliable.

Particularly in designs where a single primary sampling unit (PSU) is sampled in some strata, direct variance estimates for these strata are not possible. Alternative variance estimation procedures for strata with one PSU sampled are studied. The first option is a collapsed strata variance estimator followed by synthetic variance redistribution to the stratum level. The second alternative is the use of generalized variance functions (GVFs) estimated by direct variance estimates in strata with more than one PSU sampled for predicting variances in strata with only one PSU in the sample. The GVF methodology shows advantages in simulation studies and in an application of a stratified multi-stage sample survey conducted by Iowa's State Board of Education (ISBE).

In the context of small area estimation, hierarchical Bayesian (HB) analysis is proposed to produce more reliable estimates of small area quantities than direct estimation. A method that benchmarks the HB estimates to the higher level direct estimates and measures the relative inflation of posterior mean squared error in the posterior predictions is developed to evaluate the performance of hierarchical models. Both numerical and graphical summaries of the posterior predictive discrepancy measures are available. The benchmarked HB posterior predictive model comparison method is shown to be able to select proper models effectively in an illustrative example. The method is then applied to fitting models to the ISBE survey data. 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. The proposed method is used to select a generalized linear mixed model for analyzing the data. Potential applications extend beyond the survey and education contexts.

CHAPTER 1. Introduction

Complex surveys are commonly used for gathering information from finite populations in studies in social science, public health, education and numerous other public issues. The surveys often involve stratification and clustering structures and sample units are drawn at multiple stages. Many surveys, however, are subject to various restrictions on budget, time and resources. Given these restrictions, the surveys are usually designed to produce specific accuracy of direct estimation at high levels of aggregation such as regions or divisions in national surveys. Consequently, sample sizes for small geographical areas or small subpopulations are typically very small and the direct survey estimates for these areas are very unreliable. When statistics for small areas are of interest, which is often the case in reality, methodologies that can improve on the direct estimation and produce more accurate and precise estimates of small area statistics and better assessments of the precision of the estimates are demanded.

A survey on transcripts of Iowa's public high school students served as a motivating example and is described in Chapter 2. The survey was conducted by Iowa State Board of Education (ISBE). It was a stratified multi-stage survey to study the employment-related courses taken by Iowa's public high school students. A very limited sample was drawn from a two-way stratification cross classified by district size and area education agency (AEA). As a result, only one or two primary sampling units were drawn from each stratum, which served as a "small area" in the survey. The survey posed two interrelated challenges. One is to provide reliable estimates of the small area means. The other is to reasonably describe the precision of the estimates. We will tackle these problems within the framework of design-based estimation and model-based inference separately.

In design-based estimation, first we need to choose an estimator that can achieve more

precision subject to the limited sample size. A ratio estimator that utilizes an auxiliary variable which is related to the outcome variable in the survey and usually produces smaller variance is considered. Second, in the design where only a single primary sampling unit is sampled in a certain stratum, there is not enough degrees of freedom to estimate the variance for the estimate for a one-PSU stratum. We need to find ways to quantify the variation of the estimates. This is even more difficult when variance estimates for individual strata are needed instead of for a group of strata.

Traditional applications employ the collapsed strata estimator. The method collapses strata into groups, treats the strata within a group as independent sample units and computes the variance of the collapsed group to approximate the sum of the variances of the strata in the group. The method has been used widely in survey applications. However, the method is not developed for producing variance estimates for individual strata. In our application, we propose new adapted methods in Chapter 3 for estimating variances of one-PSU strata. The first method uses the collapsed strata variance estimator followed by synthetic variance redistribution. The redistribution method is suggested based on characteristics of our particular target population and sample design. To be applied to other surveys, the redistribution approach should be reexamined and tailored to the specific application. The second method uses a restricted generalized variance function (RGVF) with modifications for the one-PSU-per-stratum design. The restriction is imposed for avoiding negative variance prediction. One modifications is proposed to simply adjust the coefficients of the fitted RGVF for differences in sample sizes. A different modification collapses the one-PSU strata into groups, which have the same design as the strata used to fit the RGVF in terms of the sample size of PSUs, and predicts the variance of the collapsed group followed by synthetic variance redistribution. The performance of the proposed methods will be examined and compared through simulations in simple general settings and in the ISBE application.

As mentioned earlier, the design-based estimation is subject to the curse of the small sample size.. We then consider an alternative method that does not rely on the asymptotic approximation for large sample size and can be applied to both large and small sample cases. We choose

to use a hierarchical Bayesian method which provides a unified framework for dealing with problems with large and small sample sizes. The method allows the use of hierarchical models to better characterize complex data structures, such as cross-sectional effects, geographical dependence and time series correlation, and relation with auxiliary variables. The estimation procedure is based on approximating the posterior distributions using Markov Chain Monte Carlo (MCMC) methods, which are easy to implement with the aid of the advanced computing capacity and well-developed computing techniques.

In the hierarchical Bayesian analysis for the ISBE survey data, we propose two families of generalized linear mixed models (GLMMs) to characterize the correlation structure of the data in Chapter 4. The hierarchical Bayesian (HB) estimators for estimating the stratum means and the variances of the HB estimates are derived under the proposed GLMMs. The HB estimates of the stratum means and variances can be obtained using the iterative simulates from the approximate posterior distribution based on MCMC simulation. We investigate the precision of the HB estimator relative to ratio estimator based on an illustration example using a simulated population. Then we will conduct an exploratory data analysis for the actual ISBE survey data using an informal model building process and the proposed two families of GLMMs. The models were built in an iterative procedure. A simple model is initiated. In each cycle, the current model was fitted to the data and checked for possible deficiency and the next model is proposed based on the model checking result of the current model. The process continues until a succinct model that characterizes data features properly and adequately is chosen. The results of data analysis based on the examined models are discussed within the context of the survey data.

A crucial issue in HB analysis is selecting an appropriate model to analyze the data. Many practices in model diagnostics and model evaluation have been done informally in a variety of applications. However, formal methods of model checking and comparison would better help the practitioners in analyzing the data. In Chapter 5, we will focus on study of formal methods for model comparison and model selection in Bayesian framework. Some existing methods from the posterior predictive perspective, such as the posterior predictive p-value,

the L-criterion, the deviance information criterion, are reviewed. We then propose a novel method which uses a new discrepancy that measures the inflation of posterior mean squared error due to benchmarking the HB estimator to the reliable direct estimator in large regions. The new measure can be used in the traditional posterior predictive checking. We also suggest using a method of graphical examination in the posterior predictive checking which can yield added insight into model comparison. Our method is examined and compared with the existing methods discussed above in an illustrative example and then in the analysis of the actual ISBE survey data. The administrative variables from the Common Core Data (CCD) from the National Center for Education Statistics were examined and used as covariate variables in the models in the real data analysis. The proposed benchmarked hierarchical Bayesian posterior predictive model comparison method should be able to be adapted and applied to applications beyond survey and education context such as disease mapping and statistical quality control.

In the following chapters, we will begin by discussing the ISBE transcript survey and some related design and estimation issues in Chapter 2. Then we will address the issue of variance estimation in designs with a single PSU per stratum in Chapter 3. In Chapter 4, we will conduct a hierarchical Bayesian analysis of the ISBE survey data using some GLMMs proposed based on the data structure. In Chapter 5, we will develop a novel method of hierarchical Bayesian model selection by using benchmarking in defining a discrepancy measure in posterior predictive checking. Chapter 6 will have some conclusions and discussions of possible future study.

CHAPTER 2. Transcript Survey Study by Iowa's State Board of Education

Education has always a topic of deep concern to the whole society. Teaching and tutoring young people and help them better prepare for their future lives is a major task for every educational institution. The high schools which play an very important role in education of young people should think much of the question especially. Students may have different life choices after graduating from high school. Some students go to the university or college, whereas others start work or follow other paths. It is challenging for high schools to prepare their students for all of these possible future paths. Also high schools differ from each other in many aspects, such as locations (urban or rural), administrative policies, budgets, as well as the enrollment sizes and course offerings. It is very likely that some schools are of high quality and some schools may outperform others in certain aspects. Due to these concerns, Iowa's State Board of Education (ISBE) is especially interested in studying factors that are related to performance of high schools in terms of helping students formulate their life choices and obtain the capability to pursue their choices. A further goal is to set policies that help under achieving schools improve their performance.

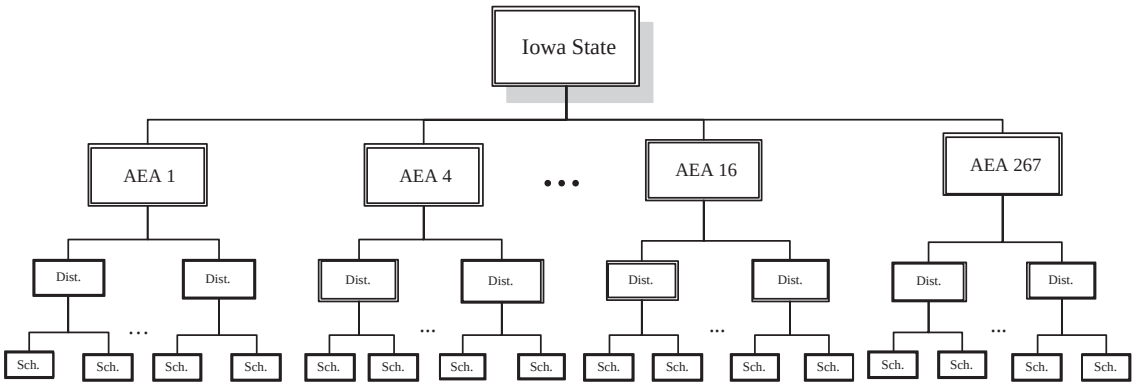
For seeking the answers to the above questions, ISBE planned a series of surveys of high school seniors from more than 300 public high schools in Iowa. Some of these surveys were to study the transcripts of high school students to see what classes are offered in different schools and what classes the students actually take. The courses usually offered by public high schools include college preparation, employment preparation, and remedial courses. Students in a public high school are enrolled in general education or special education depending on if individual educational plan (IEP) for physical, mental, or behavioral disabilities is taken.

Students having special education are categorized into three levels, and students in the first two levels often have transcripts that look like those of students having general education. Students in the third level group who do not have comparable transcripts are not included in the study. Private schools and alternative schools are also excluded from the study population.

In 2004, representatives of ISBE approached the Center for Survey Statistics and Methodology (CSSM) at Iowa State University (ISU) for help in planning these surveys. The purpose of one of the surveys was 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. EP courses belong to a diverse set of courses including those in information technology, accounting and business, trades and professions, and agricultural management (Bradby et al. 1995). A primary concern of the survey was 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.

2.1 Sample Design

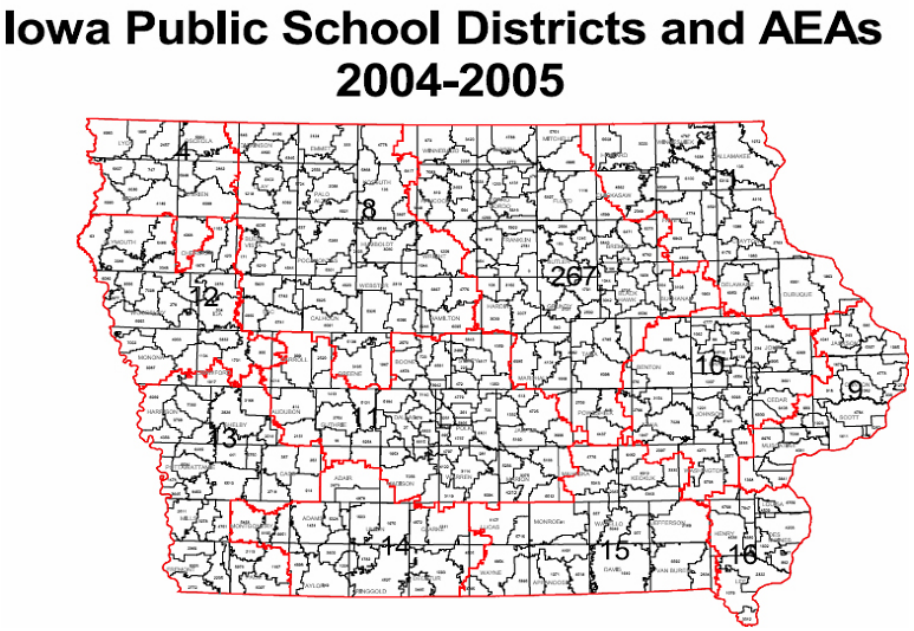
Figure 2.1 The structure of the public educational system in Iowa.



As shown in Figure 2.1, the public educational system in Iowa is organized into schools, districts, and Area Education Agencies (AEAs). There are about 340 public high schools in Iowa. For the purpose of administration and organization, the schools are grouped together into districts for setting policy and determination of curriculum. Districts are then divided into twelve administrative regions overseen by twelve AEAs. The geographical division of the

twelve AEAs is shown in Figure 2.2. The AEAs are not numbered consecutively. The AEAs with consecutive numbers are not necessary geographically adjacent. Some AEAs have larger acreage coverage such as AEAs 8, 11, and 267. Some of them are geographically small such as AEAs 4, 9, and 16. The school districts in Iowa are also categorized into three size levels due to their very different enrollment sizes. Small school districts have less than 250 students. Medium districts have 250 to less than 2,500 students. Large districts have 2,500 or more students.

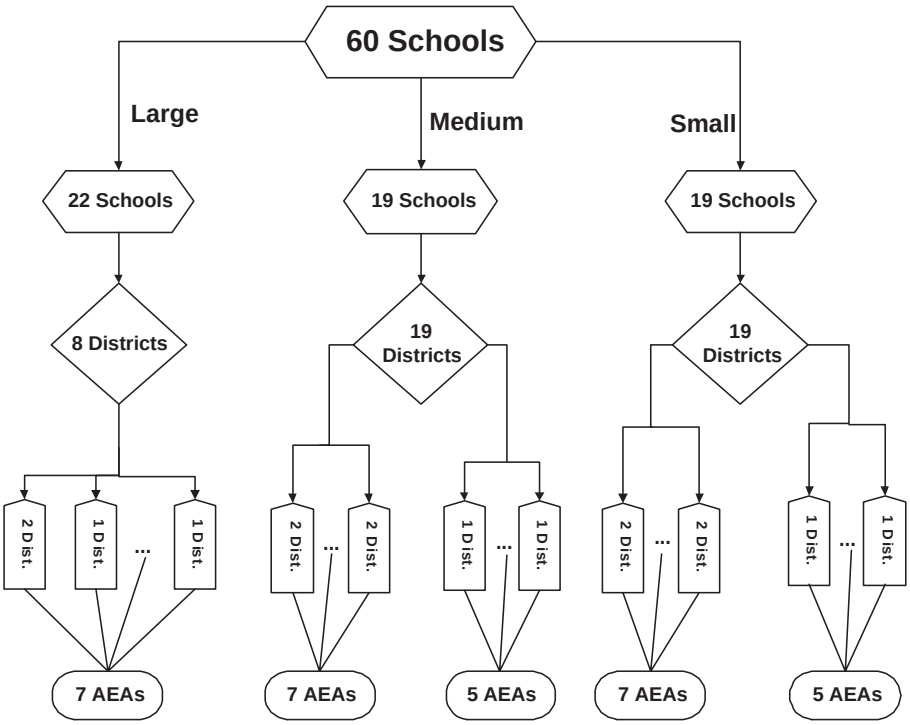
Figure 2.2 The geographical division of 12 AEAs in Iowa.



The sample survey was designed to produce estimates of the average numbers of EP courses taken by high school students for the State of Iowa and populations of small, medium, and large school districts. Considering the objective of the survey, the structure of the target population and the resource restrictions, a stratified multi-stage survey was planned. District size and AEA were used as stratifying variables. Because five AEAs have no large school districts, there are $12 \times 3 - 5 = 31$ strata in total. At the first stage of sampling, school districts were sampled within each stratum. All large school districts were taken with certainty due to their extreme sizes. Districts of medium or small size were sampled with probability proportional to size sampling without replacement. The total enrollment of a district was used as the measure

of size. At the second stage, for political reasons, all schools within sampled districts were included in data collection. Only large and one medium school district in Iowa have more than one high school. At the third stage of sampling, a simple random sample of students was selected in each sampled school. The samples were split between grade nine and grade twelve students having general and special education. The sample design was actually a two-stage design since all schools in selected districts were sampled with certainty.

Figure 2.3 The sample allocation in the design of 60 sample schools.



Due to the limited resources, CSSM was informed by ISBE that the actual survey could only collect data from 60 schools and no more than 12,000 students. In the design with 60 sample schools, 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. A flowchart that explicates the sample allocation is displayed in Figure 2.3. The numbers of school districts in the sample design and in the population within

each stratum are shown in Table 2.1, which also present the sampling fraction of PSUs in each stratum indirectly.

Table 2.1 The numbers of school districts within each stratum in the sample design and in the population. In each pair of parentheses, the number in front of the semicolon is the number of school districts in the population of the stratum, and the number after the semicolon is the number of districts in the sample.

	Large	Medium	Small
AEA1	(1;1)	(12;2)	(9;1)
AEA4	(0;0)	(6;1)	(7;1)
AEA8	(0;0)	(16;2)	(27;2)
AEA9	(1;1)	(13;2)	(7;1)
AEA10	(2;2)	(15;2)	(14;2)
AEA11	(1;1)	(29;2)	(22;2)
AEA12	(1;1)	(6;1)	(13;2)
AEA13	(1;1)	(7;2)	(21;1)
AEA14	(0;0)	(4;1)	(13;2)
AEA15	(0;0)	(6;1)	(5;1)
AEA16	(0;0)	(6;1)	(5;1)
AEA267	(1;1)	(23;2)	(34;2)

Within each sampled school, students from ninth and twelfth grades having general and special education were selected by simple random sampling without replacement. To sample a total of 12,000 students in 60 schools, 50 students on average were expected to be selected from each group within each sampled school. However, the number of students in different groups in different schools are very varied. Large and medium schools usually have more than 50 students having general education in grade 9 or grade 12. But many small schools have fewer than 50 students in these groups. Also most schools had fewer than 50 IEP students total in grades 9 and 12. Many schools, in fact, had fewer than 50 students together in grades 9 and 12. In this case, a simple design was to draw a simple random sample from each group with a maximum of 50 students. This was easy to implement but would waste a lot of survey resources that were in fact available for use. A more complex design that allowed redistribution

of resources would make better use of all available resources.

Specifically, if the number of students in a group was substantially larger than 50, then initially 50 students were selected from the group. Then additional student transcripts were examined in some groups in some schools up to 200 students total in every selected school. Further, resources from schools with fewer than 200 students were assigned to other larger schools so that excess sample was collected in some bigger schools. This was possible, because it usually was feasible to do more data collection in large schools. This process of redistributing survey resources can be applied to other designs with a different number of schools and students.

To implement the sampling processes with and without resource redistribution, a computer program was developed in the R statistical computer language. For any desired number of sample schools and students, the program produces a sample specification indicating the number of students to be selected from each group within each sampled school. This program was used in repeated simulations of the sampling scheme applied to the artificial population data.

The actual data collection was proceeded by sending two data recorders to each district in the sample to examine the district course catalog and identify courses of interest, and also examine the transcript of each student in the sample and record which courses of interest were taken by the student and when they were taken. Since districts can independently decide their course names and numbers, examination of district course catalog and identification of courses or interest was a critical step in data collection.

2.2 Design and Estimation Options in the Survey

2.2.1 Direct Estimators

Two kinds of direct estimators were proposed to estimate the total (and mean) number of employment preparation courses taken by high school students in a stratum. The schools in districts of a particular size within an AEA were the members of a stratum. The first estimator used in direct survey estimation was the π-*expansion estimator* of Horvitz-Thompson (Horvitz and Thompson (1952)). Suppose a sample of n units is selected without replacement. The

Horvitz-Thompson (HT) estimator of the population total is

$$\hat{Y}_{HT} = \sum_{i=1}^{n} \frac{y_i}{\pi_i}, \tag{2.1}$$

where y_i is the observed value for the sample unit i and π_i is the inclusion probability for unit i (Cochran (1977), pages 259-261). In a multi-stage sampling design, the inclusion probability is a product of probabilities of selection at all stages. In the EP survey, within strata defined by district size and AEA, the probability of selecting a student was the product of the probabilities of selecting a district, a school within a district, and the student within a school. In a multi-phase sampling, which is relevant for the non-invariant designs that will be discussed in Section 2.2.2, an adjusted HT estimator, called the $\pi^\star$ *estimator*, in the terminology of Särndal et al. (1992) (page 347), is suggested. In a two-phase sampling, π_i in (1) is replaced by

$$\pi_i^\star = \pi_{ai}\pi_{i|S_a}, \tag{2.2}$$

where a represents the first phase, π_{ai} denotes the probability of selection of unit i in the first phase sample, and $\pi_{i|S_a}$ is the probability of selection of unit i given a sample S_a realized in the first phase.

The second design-based estimator used in our application was the *ratio estimator* (Cochran (1977), chapter 6). Ratio estimation works well when a convenient and inexpensive auxiliary variable that is correlated with the response variable is available for all units in the population. In the ISBE survey, the enrollment was known for all schools and was suspected to be positively related with the number of enrollments in EP courses in the school. So it could be used as an auxiliary variable in ratio estimation. Let X denotes the total of the auxiliary variable which is approximately linearly related with the outcome variable whose total is denoted by Y. In the EP survey, Y was the total number of EP courses taken by twelfth grade students in a stratum, and X was the enrollment of twelfth grades students in the stratum. The ratio estimator of Y is given by

$$\hat{Y}_{RA} = X\frac{\hat{Y}}{\hat{X}}, \tag{2.3}$$

where $\hat{Y}$ and $\hat{X}$ could be the Horvitz-Thompson estimators of Y and X. The ratio estimator is not unbiased, but it is design consistent. It is usually used to produce smaller variance than

HT estimator by utilizing auxiliary information. It is very precise when the population points (y_i, x_i) are tightly scattered around a straight line through the origin.

The estimates of totals in the whole state, size levels, and AEAs are the sum of estimates of totals in all strata contained in those aggregations. The estimates of means are the estimates of totals divided by the number of students in the relevant aggregation.

Table 2.2 The standard deviation (SD) and the root of mean squared error (RMSE) of total estimates using HT and ratio estimators for three aggregations.

	SD		RMSE	
Aggregation	HT	Ratio	HT	Ratio
State	5510	4914	7398	6567
Medium	5357	4789	7189	6422
Small	1207	1003	1690	1400

The performance of the HT and ratio estimators were examined using a simulated population through a Monte Carlo study. A population database of twelfth grade students was created through simulation. The numbers of EP courses taken by students in a school were generated as independent Poisson random variables with a rate for the school. The Poisson rates were generated independently from a random effects model with main effects due to school size and AEA, which are the factors used for the actual stratification. The simulations did a reasonable job of creating a population database not unlike the preliminary data, in terms of number of courses taken by students.

Table 2.2 shows the standard deviation and the root of mean squared error of total estimates using HT and ratio estimators for the whole state and the medium and small size levels. Since all schools in large districts were taken with certainty, the estimate of total (and mean) and the variance was trivial. The ratio estimator was not applied at the large size level. The ratio estimator outperforms the HT estimator in terms of producing smaller variance and mean squared error in the simulation. Therefore, when the number of EP courses is actually related to the school enrollment, using a ratio estimator by utilizing enrollment as the auxiliary

variable would improve the precision of estimation over the Horvitz-Thompson estimator.

2.2.2 An Invariant And A Non-invariant Design

The standard stratified multi-stage design is invariant in the sense that the same subsample design for a PSU is used every time the PSU is included in a first stage selection. If up to 50 students from each of the four groups in a selected school are sampled independently by simple random sampling with no redistribution in the case of small schools, then the design is a stratified multi-stage design and the standard formulas for estimators of means and totals and estimators of variances are applicable. The invariant design is easy to operate. The realized overall number of students in the sample, however, is not fixed and would be less than the specified maximum. This will increase the variance contributed by the terminal stage of sampling especially for large schools.

On the other hand, if excess sample is redistributed across districts and schools, the inclusion of certain districts or schools in the sample affects student selection probabilities in other clusters. In the terminology of Särndal et al. (1992) (page 134), the design is not invariant. The non-invariant multi-stage design can be thought of as a multi-phase sample design. In a multi-phase sample design, the subsample design depends on the entire selected first phase sample. In the EP survey, if further resource redistribution is planned, then the design is a stratified multi-phase design and the standard formulas (see Särndal et al. (1992), chapter 9) for estimating totals and variances of total estimates for a multi-phase design should be employed. The estimators of both totals and variances are unbiased. Although the design is more complicated to implement and estimation formulas are more involved, the non-invariant multi-phase design makes use of all available resources and tends to decrease the terminal phase variation due to a larger sample in that phase.

Table 2.3 shows the estimates of coefficient of variation (CVEs) of the HT and ratio estimators under the invariant and non-invariant designs for the whole state and the medium and small size levels. The variance estimates for the medium and small size levels were obtained using the collapsed strata estimator. The non-invariant design resulted in smaller CVEs using

Table 2.3 CVEs (displayed in %) of Horvitz-Thompson and ratio estimators under invariant and non-invariant designs for three aggregations.

Estimator	Invariant Design	Non-Invariant Design
HT		
State	3.60	3.55
Medium	6.46	6.38
Small	6.29	6.21
Ratio		
State	3.42	3.34
Medium	6.19	6.05
Small	5.27	5.18

either the HT or the ratio estimator. Also, the ratio estimator produced consistently smaller CVEs than the HT estimator under the invariant and non-invariant designs. Therefore, if it is possible to implement, the non-invariant design has some definite advantages.

2.2.3 A Cost Evaluation

Iowa's State Board of Education decided on 60 schools instead of 70 as was recommended, because it was operationally feasible in terms of budget, staff, and coordination with schools. However, in our simulation study, we examined the case of sampling 70 schools (two PSUs per stratum for all strata) and assuming fewer students per sampled school. The 70 school design had lower variances at most levels of aggregation beyond the schools and districts. Therefore, we further considered investigating the implied costs of adding schools to the sample and hoped to quantify the implied trade-off between cost and variance.

Costs in this school board survey generally come from general administration, data processing, sampling districts, and sampling students. Since additional schools will be from the small and medium districts, each additional district adds one school. In order to examine this issue, it was assumed that all sampled students cost the same in terms of data collection and processing and each sampled student costs one cost unit. Also assumed that all sampled schools cost the same. Supposed that an additional school "cost" a cost units each. That is, in order

to spend resources to code a new course catalog and to interact with school administrators, *a* fewer students across the whole study had transcripts reviewed. Considering that the between schools variation played a significant role in the variance of total estimates, it was of interest to study with a fixed overall budget how much one could benefit from increasing the number of sample schools while correspondingly reducing the number of sample students overall and per school.

Table 2.4 Empirical percentiles over 1,000 simulations of the relative decrease in variance estimates due to adding more schools to the sample with fixed total cost per school. Decrease is relative to a sample with 60 schools. Variance estimates for the cases of 60 and 65 sample schools used the collapsed strata estimator. Cost factor per school is the reduction in the total number of students in the sample per additional school. IQR is the inter-quantile range of variance reductions.

Number of Schools		Cost Factor Per School 200	150	100	50	0
Students	65	11000	11250	11500	11750	12000
Median		0.118	0.108	0.112	0.105	0.101
IQR		0.315	0.326	0.336	0.344	0.342
Students	70	10000	10500	11000	11500	12000
Median		0.169	0.164	0.160	0.158	0.141
IQR		0.338	0.323	0.308	0.331	0.317
Students	75	9000	9750	10500	11250	12000
Median		0.202	0.212	0.193	0.201	0.196
IQR		0.230	0.213	0.220	0.221	0.236
Students	80	8000	9000	10000	11000	12000
Median		0.238	0.242	0.253	0.252	0.255
IQR		0.192	0.196	0.188	0.189	0.176

Based on the numerical results in Table 2.4, if one samples 65 schools, the variance estimates were reduced about ten percent on average. Larger costs per school, at least in the range of costs and numbers of districts considered, did not seem to make much of a difference; the number of sample schools was much more influential. After 65 schools, variance estimates decreased about five percent for every five schools added. Besides the decrease in variance, the interquartile range of variance reductions over 1,000 simulations also decreased as the number

of schools increases to 75 or 80 schools. This means that there was less variation in variance reductions, which means that variances were estimated with more stability. Therefore, it seems that even accounting for higher costs, adding more schools to the survey would produce better estimates of total and mean and better estimates of variance.

2.3 Summary

A survey for studying the transcripts of Iowa public high school students was planned by ISBE. A stratified multi-stage design was proposed to produce the estimates of the average numbers of EP courses taken by students for the whole state and populations of large, medium and small school districts. The estimates for twelve AEAs and individual strata cross classified by district size and AEAs are of interest as well. All large school districts were taken with certainty. Medium or small districts within a stratum were sampled with probability proportional to size without replacement. A simple random sample of students having general and special education from grade 9 and grade 12 was drawn within each sampled school. Further restrictions on the number of sampled schools and students were required. A redistribution of excess students from small groups to large groups within sample schools and further from small schools to large schools was suggested in order to make use of more available resources.

Some design and estimation options were examined through simulation in a preliminary study. In a situation that an auxiliary variable which is believed to be related to the response variable is available, a ratio estimator by utilizing the auxiliary variable is often suggested to improve the precision of estimation. In the ISBE application, the outcome in the survey which is the number of EP courses taken by students was considered related to the enrollment in a school level. The ratio estimator was shown to produce smaller variance and mean squared error in a Monte Carlo study based on a simulated population.

In the situation that extra resources can be redistributed across sample clusters, a non-invariant stratified multi-phase design can be applied. Compared with an invariant stratified multi-stage design, the non-invariant design makes full use of available resources and reduces the bias and variance of the estimator at a cost of more complicated implementation. The

benefit of using a non-invariant design in the ISBE survey was proved by the reduced variances of the total (and mean) estimates at various levels of aggregations of strata.

The numerical results in the simulation indicated that the between cluster variation might be more influential than the within cluster variation. To reduce the overall variation, it would be a good idea to increase the number of clusters rather than sampling a greater number of units within clusters. With fixed overall cost, the impact on variance estimates by increasing the number of sample schools while assuming the cost of sampling a district to be in a reasonable interval was studied. The number of sample schools showed a definite effect on the variance estimate. The cost of sampling a district, in the range of alternatives considered, only affected the variance estimate on a small scale. If practically possible, it would be strongly advisable to increase the number of schools in the sample.

Due to the budget, time and resource restrictions, the survey could sample a very limited number of schools. As a result, only a single school district could be sampled from some strata. In a situation that variance estimates are needed for these strata, in which there are not enough degrees of freedom to make direct estimation, alternative methods to produce reasonable variance estimates for one-PSU-per-stratum strata are needed to be studied. We will address this issue in Chapter 3. Further, the direct estimates for individual strata in the ISBE survey would be generally not reliable due to the very small sample sizes (one or two PSUs). Alternative estimators that make explicit use of hierarchical models to "borrow strength" cross related school districts and produce more reliable estimates of strata with small sample sizes will be studied in Chapter 4.

CHAPTER 3. Variance Estimation in a One-per-stratum Design

The sample design in the ISBE survey posed a very challenging problem when trying to estimate variances of estimates for individual strata. Due to the budget restriction, the survey could sample only 60 schools from the whole state. After taking all large schools with certainty, in each of the medium and small size levels, 14 schools were taken from seven AEAs that had more school districts and 5 schools were taken from the remaining five AEAs with fewer school districts. The existence of the strata with a single PSU sampled makes it difficult for variance estimation in these strata as well as AEAs that contain these strata.

Since some strata consisting of medium and small school districts had one single PSU selected within the stratum, there were not enough degrees of freedom to directly estimate the stratum variance. Standard approaches, such as the collapsed strata variance estimator (Cochran (1977), section 5A.12) consider estimation of variances for aggregations of strata, but not for individual strata. The focus of this chapter is on variance estimation in the situation of one-per-stratum sampling. Specifically, it concerns estimates of variances for individual strata with only one unit sampled. New adaptations of two procedures are compared. The first is variance estimation using collapsed strata variance estimators followed by synthetic variance redistribution. The second is to use restricted generalized variance functions with modifications for designs with one primary sampling unit per stratum.

Section 3.1 discusses some of the existing approaches for variance estimation in one-per-stratum designs. Section 3.2 discusses the proposed methods for estimating variances of individual strata with a single PSU in the sample. Section 3.3 describes results of two simulation studies. Section 3.4 contains results for the ISBE EP survey based on simulation. Section 3.5 presents summary and discussion.

3.1 Existing approaches for variance estimation in One-Per-Stratum Designs

Imagine that one cluster or primary sampling unit is selected in a stratum. Assume within the cluster that units are taken by simple random sampling. If not much difference among clusters exists within the stratum, then a reasonable simple approximation to the variance of the total estimate within the stratum would be the variance of a simple random sample (SRS) from the population in the stratum. Similarly, using a difference estimator as described in Wolter (1985) (chapter 7) for variance estimation in systematic sampling should work well if population units are randomly associated into clusters. If there is strong or even moderate homogeneity within clusters, however, then these estimators will (significantly) underestimate the true variance.

The collapsed strata estimator (Cochran (1977), section 5A.12) is a well-known estimator for variance estimation in the one-per-stratum problem (see also remark 3.7.1 in Särndal et al. (1992)). The procedure collapses strata with one unit per stratum into groups and treats the strata in a group as independent samples from the combined stratum. The collapsed strata estimator usually overestimates the group variance, but the overestimation can be controlled by grouping strata together that are as similar as possible in terms of the stratum characteristic being measured and not too different in stratum size. There is also a risk of serious understatement of variance if strata are grouped after observing the sample. Consequently, strata groups should be formed before sampling.

Hansen et al. (1953) proposed a modified collapsed strata estimator using auxiliary variables associated with strata totals in variance estimation. Hartley et al. (1969) proposed an estimation method using one or two auxiliary variables that are correlated with the strata means, which might lead to smaller bias in variance estimation in many situations. Shapiro et al. (1980) recommended a without replacement variance estimator which outperforms the collapsed strata estimator in terms of both smaller bias and smaller variance.

Isaki and Fuller (1982) proposed a regression predictor by constructing sample designs and estimators under a linear regression superpopulation model. The evaluation of estimators

is based on the anticipated variance, which is the variance under the sample design and the superpopulation model. Isaki (1983) compared variance estimation using auxiliary information under some commonly used sample designs and under the regression prediction method through a Monte Carlo study and demonstrated an improvement in terms of bias and mean squared error. Zhao et al. (1991) proposed a variance estimator of the generalized regression estimator and claimed that it has better performance than two alternatives proposed by Särndal et al. (1989).

3.2 Adapted Methods for Variance Estimation in One-Per-Stratum Strata

This section proposes two new methods for estimating stratum variances. The first method employs proportional redistribution in addition to collapsed strata variance estimation. The second method uses a restricted generalized variance function (RGVF) to produce non-negative predictions of stratum variances and makes modifications due to sample size effects.

3.2.1 Collapsing strata synthetic estimation of stratum variances

In the ISBE survey, collapsing can be implemented separately among the strata containing small and medium sized districts with one district in the sample. First arrange the strata in a non-increasing sequence based on the total enrollment size. Then collapse strata into pairs or groups sequentially. The variance of a group consisting of L_g strata can be estimated by

$$\hat{V}_{coll}\left(\hat{t}^{(g)}\right) = \frac{L_g}{L_g - 1} \sum_{k=1}^{L_g} \left(\hat{t}_k^{(g)} - \frac{\sum_{k=1}^{L_g} \hat{t}_k^{(g)}}{L_g} \right)^2 \tag{3.1}$$

where $\hat{t}_k^{(g)}$ is the total estimate of the k^{th} stratum in the group and $\hat{t}^{(g)}$ is the total estimate of the collapsed group: $\hat{t}^{(g)} = \sum_{k=1}^{L_g} \hat{t}_k^{(g)}$.

To produce variance estimates of individual strata within the group, one simple method is to use the variance estimate of the collapsed group as the estimates for individual strata, which obviously will almost always overestimate the variances for individual strata. In our application, we propose *proportional redistribution* of variance based on squared total enrollment size times the within stratum variation of units. The reason for this redistribution is that

estimates of strata within the group are independent, and thus the variance of the sum of estimates of strata equals the sum of variances of estimates of strata. Since the sample sizes in the strata within the same group are quite similar in the survey, when the intracluster correlations are close for strata in the same group, the ratio of variances of two strata within the group is approximately equal to the ratio of the products of squared total enrollment size and the within stratum variation of units. A within stratum variance of units can be estimated by the variance of sample units within the stratum. Then the variance of each stratum is a portion of the variance of the group with a weight proportional to the product of squared enrollment size and the within stratum variance of sample units. In a special case that the strata within the same group are homogeneous in terms of within stratum variation of units, the ratio of variances of two strata within the group is approximately equal to the ratio of squared total enrollment sizes. This redistribution, although not a standard practice, is important in this application for producing estimates of variances for individual strata as well as AEAs. The method can be referred to as *collapsing strata synthetic variance (CSSV)* estimation.

3.2.2 Modeling and generalized variance functions

The standard design-based variance estimator is usually relatively unstable with small sample size. In such a circumstance, one may consider using alternative variance estimators based on models of variances. Valliant (1987) found that in some circumstances some generalized variance functions (GVFs) could be simple to compute, approximately unbiased, reasonable in coverage levels when forming confidence intervals, and more stable than direct estimators.

The traditional reason for using GVFs is to produce variance estimates for a large number of survey statistics conveniently (Wolter (1985)). The basic idea is to model the relationship between relative variances and expectation of total estimators for a group of survey statistics that follow a common model by using the direct estimates of some members of the group and then predict the variances of other members from the estimated totals through the estimated function. In our case with one-per-stratum samples, we propose to fit the generalized variance function to direct estimates of variances for strata with two (or more) PSUs and then predict

variances for the strata with one PSU sampled.

Based on a preliminary examination of previous data, we assume the numbers of EP courses taken by high school students to be generated from a product of Poisson distributions. Then it is natural to consider a traditional GVF:

$$V_{\hat{T}}^2 = \alpha + \frac{\beta}{T}, \tag{3.2}$$

where T is the expectation of the estimator of the total of the population and $V_{\hat{T}}^2$ is the relative variance or rel-variance of the total estimator $\hat{T}$, which is defined as

$$V_{\hat{T}}^2 = V(\hat{T})/T^2. \tag{3.3}$$

The $V(\hat{T})$ is the variance of the total estimator $\hat{T}$. The unknown parameters in the GVF can be estimated by using *iteratively reweighted least squares* estimation to minimize

$$\sum \{m\left(\hat{T}; \alpha, \beta\right)\}^{-2} \{\hat{V}_{\hat{T}}^2 - m\left(\hat{T}; \alpha, \beta\right)\}^2, \tag{3.4}$$

where $V_{\hat{T}}^2 = m\left(T; \alpha, \beta\right)$ is the GVF and $\hat{V}_{\hat{T}}^2$ is the direct estimate of $V_{\hat{T}}^2$. One disadvantage of this model is that it can produce negative predictions of variance, for example, when α is negative and the total estimate is very large.

Wolter (1985) (chapter 5) suggested adding restrictions to the GVF to ensure that the function generates nonnegative predictions of variance. Let N denote the population size of the stratum. If we assume the same GVF holds for the estimator $\hat{N}$ and $V_{\hat{N}}^2 = 0$, which is equivalent to $\alpha = -\frac{\beta}{N}$, then a GVF with restrictions is given by

$$V_{\hat{T}}^2 = \beta \left(\frac{1}{T} - \frac{1}{N}\right). \tag{3.5}$$

By imposing the restrictions, the GVF can successfully avoid the negative predictions of variance. The GVF with restrictions will be referred to as the restricted GVF (RGVF).

The GVF is dependent on the finite population, the sample design and the variables of study. In applications such as the Current Population Survey (CPS) or the National Health Interview Survey (NHIS), a GVF is usually applied for a group of variables with a common or quite similar population and sample design. In our case, a GVF is proposed to be utilized

for a group of strata with similar characteristics but the subgroup of strata used for fitting the model does not have exactly the same sample design (different sample size of PSUs) as the subgroup of strata whose variances are to be predicted. Therefore, we propose to adjust the GVF method due to the difference of sample size when predicting variance. One approach is to adjust the coefficients of the estimated GVF model by the ratio of the sample sizes of designs used for fitting the model and for predicting the variance. The method could be referred to as adjusted (coefficients) GVF (AGVF) or adjusted restricted GVF (ARGVF).

An alternative method is to collapse the single PSU strata into groups such that each group has the same sample size as the strata used for producing direct variance estimates for fitting the GVF model, predict the group variance using the estimated model, and then estimate variances for individual strata by redistributing the predicted group variance to the stratum level. The redistribution could follow the same procedure proposed for CSSV estimation. The method assumes there are several strata available with the same number (greater than 1) of PSUs for fitting the GVF model. This is commonly true for highly stratified designs. Model estimates can be made separately for differently-sized groups of 1-per-stratum PSUs. If the size of 1-per-stratum PSU groups matches the size of the strata used to fit the GVF, then the GVF coefficients do not need to be adjusted. Due to the collapsing of strata, the total estimate of the collapsed group could be very big compared to total estimates for individual strata, which will increase the risk of a negative prediction of group variance. In this circumstance, the use of the restricted GVF (RGVF) is needed. The method of using a collapsing procedure combined with a generalized variance function and synthetic variance redistribution will be referred to as collapsing (strata) GVF synthetic variance (CGVFSV) estimation. When a restricted GVF is used, it is referred to collapsing (strata) restricted GVF synthetic variance (CRGVFSV) estimation.

3.3 Simulation Studies

To investigate the properties of the proposed GVF and CSSV methods, we conducted two simulation studies. The first simulation investigated the adjustment of GVF estimates for

different sample designs. Specifically, the properties of the adjusted GVF methods (AGVF, ARGVF, CRGVFSV) relative to direct variance estimation are studied. The method CGVFSV too often produces negative group variance estimates, hence is not considered further. The second simulation compares the adjusted GVF methods with CSSV estimation in the specific case of one-per-stratum designs. Different forms of GVF models including the traditional model $V^2 = \alpha + \beta/T$, the log-transformed model $\log(V^2) = \alpha + \beta \log(T)$, and the restricted model $V^2 = \beta(1/T - 1/N)$ are compared.

3.3.1 Simulation to compare direct variance estimation and GVF adjustment in general

In the first simulation study, a single finite population was simulated which consists of 10 strata each with 60 clusters and 50 units within each cluster. The element units were generated independently from Poisson distributions with stratum means systematically different across strata. That is,

$$\begin{aligned} y_{i,j,k} &\sim \text{Poisson } (\mu_i), \\ \mu_i &= 8 + 0.4i. \end{aligned}$$

Indices i, j, and k stand for the stratum, the cluster, and the unit, respectively. Two stratified two-stage sampling designs were used in which both PSUs and element units were selected using simple random sampling without replacement (SRS WOR) in each stage. In designs 1 and 2, n_1 and n_2 PSUs were selected in the first stage, respectively. Five element units were sampled from selected PSUs in the second stage in both designs. The n_1 and n_2 were assumed to be ≥ 2, so that direct variance estimates could be obtained under both designs.

The GVF model was fitted to the direct variance estimates under design 1 and used to predict the variance under design 2. Three scenarios were studied: $n_1 = 20$ and $n_2 = 5$, $n_1 = 10$ and $n_2 = 2$, and $n_1 = 4$ and $n_2 = 2$. In each case, 1,000 samples under the two stratified two-stage designs were drawn. The population total was estimated using the Horvitz-Thompson (HT) estimator. The variances of the total estimates initially were estimated using direct estimation. Then the variances under design two were estimated using the GVF methods

with and without adjustment of coefficients for sample size differences. Considering that the coefficients of GVF are usually proportional to the inverse of the sample size, we multiplied the coefficients of the estimated model by n_1/n_2 for predicting variances under design 2.

Table 3.1 The estimates of coefficient of variation (cve's) (measured in %) of total estimates using DIR (direct variance estimation), GVF (traditional model without adjustment), AGVF (GVF with adjustment of coefficients), AGVF(log) (AGVF using log-transformed model), ARGVF (restricted GVF with adjustment of coefficients), and CRGVFSV (collapsing strata restricted GVF with synthetic variance redistribution). Results are based on 1,000 replicated samples using the Horvitz-Thompson (HT) estimator.

CVEs	Variance estimation method for design 2					
Scenarios	DIR	GVF	AGVF	AGVF(log)	ARGVF	CRGVFSV
1. $n_1 = 20, n_2 = 5$	1.97	1.00	1.99	1.93	2.05	2.04
2. $n_1 = 10, n_2 = 2$	3.02	1.47	3.29	2.99	3.41	3.39
3. $n_1 = 4, n_2 = 2$	3.04	2.61	3.70	2.66	3.74	3.73

Estimates of coefficients of variation (cve's) of total estimates, using various variance estimation methods averaged over 1,000 samples are shown in Table 3.1. The cve's were calculated by $100\hat{V}^{1/2}(\hat{t})/\hat{t}$, where $\hat{t}$ and $\hat{V}(\hat{t})$ are estimates of the total and the variance of the total estimate, respectively.

In the first two scenarios, the cve's using direct variance estimation are twice that of using traditional GVF without any adjustment. In the first scenario in which the design for fitting the model has relatively large sample size ($n_1 = 20$), using GVF methods with either adjustment of coefficients or collapsing strata with synthetic variance redistribution produce cve's very close to those using direct estimation. In the second scenario with moderate sample size ($n_1 = 10$), GVF methods with adjustments produce variance estimates that are close to but a little larger than for direct estimation. In the third scenario with a very small sample size ($n_1 = 4$), GVF without adjustment generally underestimates the variance. The GVF methods with adjustments overestimate the variance a little bit more than before. The log-transformation GVF, which was pretty accurate in estimation, produces rather unstable results with such a

small sample size and extreme transformation.

The results in Table 3.1 indicate that it is necessary to make some appropriate adjustment to the model-based estimates in order to produce reasonable variance estimates. This can be done either by adjusting the coefficients or collapsing strata before estimation followed by synthetic variance redistribution. In cases with large sample sizes, the adjusted GVF methods could produce as good of variance estimates as does direct estimation on average. However, when the sample size is small, the adjusted GVF methods tend to be conservative to some degree. The smaller the sample size is, the more conservative the GVF methods tend to be.

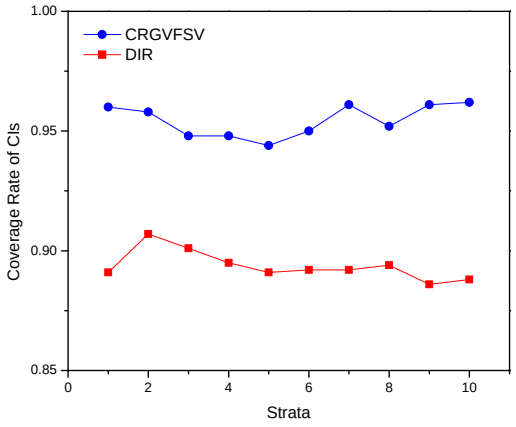


Figure 3.1 The rate of confidence intervals obtained using direct and CRGVFSV variance estimation covering the true total over 1,000 samples under the designs of $n_1 = 20$ and $n_2 = 5$.

Figure 3.1 shows the rate at which confidence intervals obtained using direct and CRGVFSV variance estimation cover the true total value over 1,000 samples under designs of $n_1 = 20$ and $n_2 = 5$. The CRGVFSV produces round 95% coverage rates for all strata. The direct estimation has consistently lower coverage rates, typically under 90%. The reason the direction estimation method has poor coverage is due to the use of normal critical value (1.96) instead of a critical value from a t distribution (2.78 with 4 degrees of freedom). If the t critical value is used, then coverage is near the nominal 95% rate.

Figure 3.2 shows the standardized root of mean squared error (SRMSE) of variance esti-

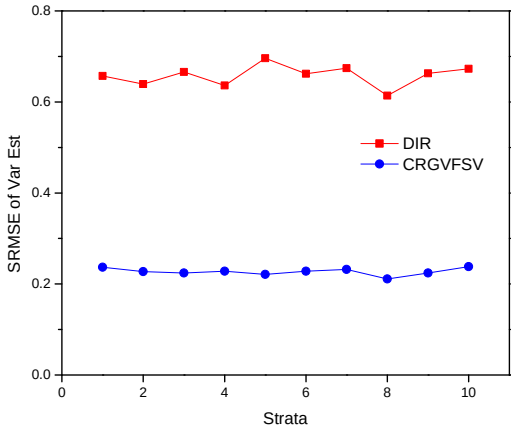


Figure 3.2 The standardized root of mean squared error (SRMSE) of variance estimates, defined in (3.6), using direct and CRGVFSV variance estimation based on 1,000 samples under the design of $n_1 = 20$ and $n_2 = 5$.

mates over $B = 1,000$ samples under the same designs. The SRMSE is defined as

$$\frac{1}{V}\sqrt{\frac{1}{B}\sum_{s=1}^{B}(\hat{V}_s - V)^2}, \tag{3.6}$$

where V is the true variance of the total estimate and $\hat{V}_s$ is the estimate of the variance of the total estimate based on the s^{th} sample. The CRGVFSV method has smaller SRMSE's than direct estimation for all strata, which indicates the former produces more precise variance estimates than the latter. The comparison of coefficient of variation (CV) of variance estimates, which is estimated by

$$\frac{1}{\hat{\bar{V}}}\sqrt{\frac{1}{(B-1)}\sum_{s=1}^{B}(\hat{V}_s - \hat{\bar{V}})^2}, \tag{3.7}$$

where $\hat{\bar{V}} = \sum_{s=1}^{B}\hat{V}_s/B$, using CRGVFSV and direct variance estimation, is not displayed here but the result shows the former produces less variation among variance estimates, which indicates more stable performance than the latter.

The AGVF and ARGVF methods perform similarly to the CRGVFSV method and show the same properties relative to direct variance estimation under the same designs. The other two scenarios with smaller sample sizes show exactly the same pattern except the benefits

of using adjusted GVF methods are more compared to the large sample cases. Therefore, despite the fact that the adjusted GVF methods could be conservative to some degree for small sample cases, the greater gains in producing higher coverage rate of confidence intervals and more precision and stability of variance estimate are consistent and appealing. In this illustrative example, the log-transformed model tends to underestimate the variance when sample size is small and hence produces lower coverage rates than the other GVF methods.

3.3.2 Simulation to study CSSV and GVF adjustment in single PSU designs

In the second simulation study, which focused on the application in one-per-stratum designs, a population was created to be composed of 50 strata each with 20 clusters ranging uniformly in size from 30 to 80 units. The units in a cluster were independently generated from a Poisson distribution with a rate for the cluster:

$$y_{h,i,j}|\lambda_{h,i} \sim \text{i.i.d. Poisson}\left(\lambda_{h,i}\right), \tag{3.8}$$

where h denotes the stratum, i indicates the cluster, and j represents the element unit. The Poisson rates were assumed to be systematically different across strata and associated with random variation for individual clusters as

$$\lambda_{h,i} = 0.1h + \tau_{h,i}, \tag{3.9}$$

where $\tau_{h,i} \sim \text{Uniform}\left(5, 10\right)$. The intracluster correlation coefficients are around 0.01 for all strata. We want to estimate the totals (and means) of individual strata as well as the variances of the estimates.

One thousand ($B = 1,000$) independent stratified two-stage samples were drawn. One or two PSUs were sampled from each stratum in the first stage using simple random sampling or probability proportional to size sampling without replacement. Within each selected PSU, five element units were sampled by simple random sampling. The Horvitz-Thompson estimator was used to estimate the totals (and means) of individual strata. The variances of strata with two PSUs sampled were estimated using direct variance estimation. To produce variance estimations for strata with a single PSU sampled, the CSSV and the adjusted GVF methods

including AGVF, ARGVF and CRGVFSV were employed. To see the effect of the degrees of freedom for fitting the GVFs, we compared designs in which either 10, 20, or 30 strata with two PSUs were selected. The designs had correspondingly 40, 30, or 20 strata with only one PSU sampled.

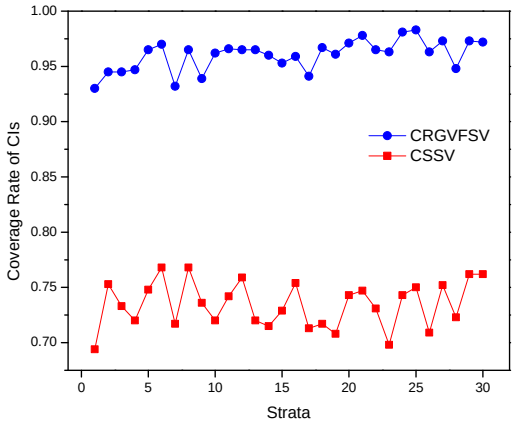


Figure 3.3 The coverage rate of confidence intervals using CSSV and CRGVFSV methods for one-PSU strata over 1,000 samples under a PPS design. Strata 1 to 30 (displayed) each have one PSU selected. Twenty strata (not shown) have two PSUs selected and are used to fit the RGVF.

We will consider the PPS design with 20 strata having two PSUs per stratum as an example to illustrate the coverage levels of confidence intervals and the stability of performance of variance estimation methods. Figure 3.3 shows the rate of confidence interval coverage of true totals for individual strata with one PSU sampled using CSSV and CRGVFSV methods. Rates are computed based on 1,000 replicate samples. The coverage rates using the CRGVFSV method are around 95%, whereas the rates using the CSSV method are between 70% and 80%. The situation using the ARGVF method is very similar to that of CRGVFSV. The coverage rates using AGVF are slightly lower than using ARGVF or CRGVFSV for most of the one-PSU strata. Still they are above 85%, which is higher than when using CSSV estimation.

Figure 3.4 displays the estimated coefficients of variation (cve's) of variance estimates using CSSV and CRGVFSV methods for individual strata with a single PSU selected over

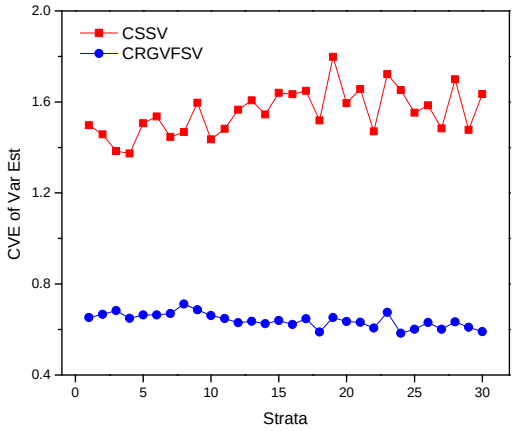


Figure 3.4 The estimated coefficients of variation (cve's) of variance estimates using CSSV and CRGVFSV methods for one-PSU strata over 1,000 samples under a PPS design. Strata 1 to 30 (displayed) each have one PSU selected. Twenty strata (not displayed) have two PSUs selected and are used to fit the RGVF.

1,000 samples. Apparently, the CRGVFSV method produces consistently smaller cve's than CSSV estimation. This comparison indicates that the CRGVFSV method is more reliable than the CSSV method. The ARGVF method performs similarly to CRGVFSV. The AGVF method, however, shows more variation in variance estimation. This high variation resulted from quite a few negative predictions of variance, which are truncated at zero, over 1,000 samples. This problem also is the reason for lower coverage rates using AGVF instead of ARGVF and CRGVFSV methods. The SRMSE's of variance estimates using CSSV, ARGVF, and CRGVFSV essentially are indistinguishable in this set of simulations. The results for a SRS design show the same properties as for a PPS design in terms of the comparison of the adjusted GVF methods relative to the CSSV method.

As to the comparison of designs with different numbers of strata having two PSUs selected, when the number of two-PSU strata increases, the adjusted GVF methods produce better variance estimates in terms of consistently higher coverage rates of confidence intervals and smaller SRMSE's and cve's of variance estimates for individual strata. This result indicates that increasing the degrees of freedom for fitting the GVF models does improve the prediction of

variances in terms of a higher coverage rates for confidence intervals, more precise estimates of variance, and more stable performance. Consequently, if practically possible, it is advisable to have more strata with two PSUs selected to better fit the model and to improve the predictions of variances.

3.4 Results for the Iowa Employment Preparation Survey

The actual survey data were not available for this analysis when the study was conducted. To study the problem of one-PSU-per-stratum variance estimation, a population database of employment preparation (EP) courses taken by twelfth grade students was created through simulation to match the expected pattern of responses in the survey. In contrast to the simulation in the previous section, schools and districts are of very different sizes and numbers across AEAs. The numbers of EP courses taken by students in a school were generated as independent Poisson random variables with a rate for the school. The rates of the Poisson distributions were generated independently from a random effects model with main effects due to school size and AEA, which are the actual factors used for the stratification. The population sizes in the simulation match the actual population sizes in Iowa's school districts in 2004. Based on examining preliminary data, the simulation did a reasonable job of creating a population database (Lu and Larsen 2006). The results presented in this paper are not actual results from the survey and should not be interpreted as characterizing schools in the State of Iowa.

Table 3.2 The estimated coefficients of variation (cve's) (measured in %) of total estimates using CSSV and three adjusted GVF methods. Results are based on 1,000 replicated samples, estimation using the ratio estimator

	Variance estimation method			
CVE of totals (%)	CSSV	AGVF	ARGVF	CRGVFSV
Aggregation				
State	3.34	3.43	3.57	3.29
Medium districts	6.06	6.18	6.45	5.96
Small districts	5.22	5.66	5.99	5.16

Table 3.2 shows the cve's of the total estimates at various levels of aggregation using the CSSV, AGVF, ARGVF and CRGVFSV variance estimation methods for the case of 60 sample schools. The ARGVF method produces a little bit more conservative variance estimates on average than the CSSV method. The CRGVFSV method produces smaller variance estimates than the CSSV method for the whole state and small and medium school districts. Since all schools in large districts are sampled, it is not necessary to use these methods for the large size category.

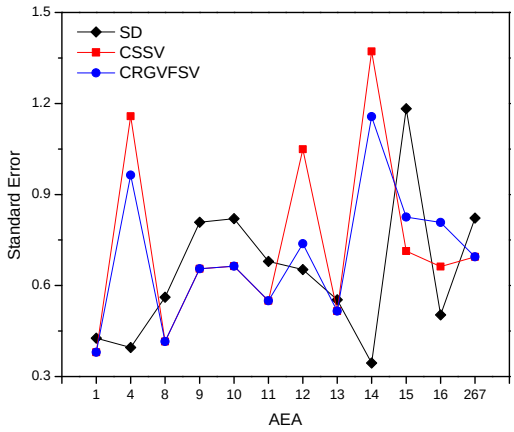


Figure 3.5 Average of standard errors of the ratio estimator using CSSV and CRGVFSV variance estimators for strata of medium districts in the case of 60 sample schools. SD is the average of the standard deviations of mean estimates over 1,000 samples.

For variance estimation in individual strata and AEAs, the CRGVFSV method still performs better than the alternatives. Figure 3.5 shows the standard errors of ratio estimates for strata of medium districts computed using CSSV and CRGVFSV variance estimation methods averaged over 1,000 samples. The five strata with one PSU per stratum in samples from the simulated population occur in AEAs 4, 12, 14, 15, and 16 (AEAs are not numbered consecutively). In four out of five simulated AEAs, the CRGVFSV method produces more precise variance estimates than CSSV estimation. Compared to CRGVFSV, ARGVF (not shown) tends to more significantly overestimate the variance. Results for strata comprised of small

districts are qualitatively the same.

Table 3.3 Number of confidence intervals obtained by using CSSV and CRGVFSV estimation out of 1,000 samples covering totals for strata with medium size districts.

Coverage	Area Education Agencies with 1 PSU				
out of 1000	4	12	14	15	16
Variance Method					
CSSV	984	864	983	707	790
AGVF	304	531	301	824	937
ARGVF	1000	983	1000	915	1000
CRGVF	1000	858	1000	808	985

Table 3.3 shows the number of confidence intervals covering the true totals using the ratio estimator and the four variance estimation methods in the five strata of medium districts with one PSU sampled. In all five strata, the confidence intervals computed by the ARGVF method have significantly higher coverage rates than those using CSSV estimation. In four strata, all except AEA 12, the CRGVFSV method produced higher coverage rates, too. The improvement in AEAs 15 and 16 are significant. The coverage rates using AGVF in AEAs 4, 12, and 14 are very low, which is due to negative predictions of variance, which are truncated at zero. Again results for strata comprised by small districts are qualitatively the same.

Figure 3.6 displays 2.5%, 25%, 50%, 75%, and 97.5% empirical percentiles of the width of confidence intervals using the CSSV and CRGVFSV methods for the ratio estimator over 1,000 simulations for the five strata of medium districts with one sampled district. In AEAs 4, 12 and 14, the CRGVFSV method produces narrower confidence intervals than CSSV estimation most of the time. In AEAs 15 and 16, even though the medians of the width of confidence intervals by the CRGVFSV method are bigger than by CSSV method, the third quantiles are smaller. The empirical ranges and inter-quartile ranges of the variance estimates are much smaller using the CRGVFSV method for all five strata. So the variance estimates produced by the CRGVFSV method are less variable than those produced by CSSV estimation, which indicates that the former has more stable performance. The ARGVF (not shown) shows a little more variation than CRGVFSV but still has similar properties relative to the CSSV method.

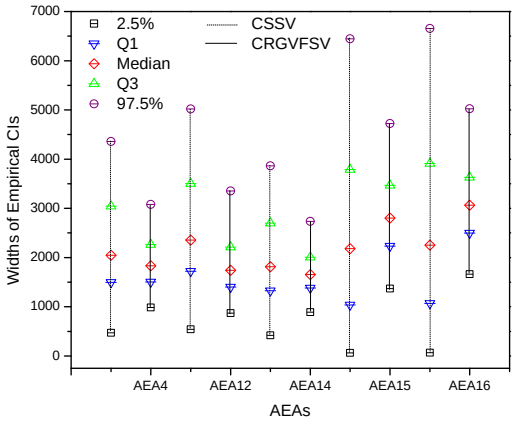


Figure 3.6 Empirical percentiles of the width of confidence intervals obtained by collapsing strata synthetic variance (CSSV) and CRGVF estimation over 1,000 simulations for strata with medium size districts using the ratio estimator.

All the results for strata containing small districts are substantially the same as those for strata containing medium districts.

We also compare the empirical percentiles of widths of confidence intervals using CSSV and AGVF methods. Due to lots of negative predictions of variance (truncated to 0 in the application), the AGVF method shows absolutely more variation among variance estimates than the CSSV method. Therefore, the potential negative prediction of variance for the traditional GVF model could cause serious problems in reality and result in poor confidence intervals. The employment of restricted GVF estimation in this situation is highly recommended.

In conclusion, numerical results show that the ARGVF and CRGVFSV methods produce better variance estimates than CSSV estimation in terms of a higher coverage rate for confidence intervals and more stable performance for individual strata as well as for the group as a whole. The ARGVF methods tend to be slightly more conservative than does CRGVFSV. The CRGVFSV method actually produces even more precise variance estimates than the CSSV method for the individual strata and at the higher levels of aggregation in the simulation. It is clear, however, that results vary more by AEA than by the method of estimation.

3.5 Summary and Discussion

A survey for which records on transcripts of Iowa public high school students served as the source of data was used to motivate the examination of variance estimation methods for designs with one-per-stratum selections of PSUs. In particular, methods for estimating variances for the strata with only one PSU were studied.

The traditional collapsing strata estimator is widely applied for estimating the variance of a total for a group of strata. In the situation that a variance estimate is needed for a stratum in which there really are not enough degrees of freedom to make a direct estimate, using a generalized variance function and choosing a reasonable model based on some model diagnostics might be possible. For example, in our application, a scatter plot of direct estimates and RGVF estimates could be helpful in checking the quality of the variance modeling.

The improvement by employing a ARGVF or CRGVFSV estimator over the direct variance estimator and over the CSSV estimator in terms of producing consistently higher coverage rates of confidence intervals and more stability of performance was demonstrated in simulation studies.

CHAPTER 4. Small Area Estimation using Hierarchical Bayesian Analysis

4.1 Small Area Estimation and Existing Methods

The term "small area" usually refers to a certain population for which reliable estimates of quantities of interest can not be obtained due to the scarcity of the available data. A "small area" could be a small geographical area such as a state, a county or a census tract. It could also be a small subpopulation such as a demographic group cross-classified by age, gender and race, or a school district. No matter how the "areas" are defined, the common features of the "small areas" are that the data could be used to estimate variables of interest for these areas are very limited and direct survey estimates are likely to have very low precision.

Small area estimation has received much attention in recent years due to the increased need for accurate and reliable descriptions of small area characteristics for many public policy issues. Small area statistics are used for apportionment of congressional seats and the allocation of government and state funds for education, public health, and numerous other expenditures. The importance of small area statistics of acceptable quality are obvious and can not be over-emphasized. However, given the constraints of limited budgets and time surveys are usually designed to ensure reliable estimates in large geographical regions or at a relatively high level of aggregation of small subgroups of a population. As a result, there are often very small sample sizes allocated to individual small areas. This will produce direct estimates with unacceptably large variances in these small areas, in which the policymakers are often interested as well.

The growing need of refined estimates of small area statistics with increased level of precision has led to extensive study and development of methods that could produce more reliable estimates of small area quantities. Due to the absence of adequate direct information, the methods are seeking to make use of information from related external sources such as various

administrative records or census data or even previous survey data through implicit or explicit models (Ghosh and Rao (1994); Rao (2003); Jiang and Lahiri (2006)).

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) estimation and hierarchical Bayes (HB) estimation have shown distinct advantages over traditional indirect estimators. Instead of using implicit models, these approaches "borrow strength" from related areas by utilizing explicit models to delineate the systematic connections among the 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.

4.2 Generalized Linear Mixed 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, two generalized linear mixed models (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 $y_{ijkl}, l = 1, \cdots, n_{ijk}$, conditionally independently follow a Poisson distribution:

$$y_{ijkl}|\lambda_{ijk} \sim \text{Poisson}\left(\omega_{ijkl}\lambda_{ijk}\right), \tag{4.1}$$

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:

$$\log\left(\lambda_{ijk}\right) = \boldsymbol{\beta}'\boldsymbol{x}_{ijk} + \tau_i + \eta_j + \zeta_{ij} + v_{ijk}. \tag{4.2}$$

The $\boldsymbol{x}_{ijk}$ 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 $\zeta_{ij} \sim N(0, \sigma_\zeta^2)$ are independent random effects from size, AEA, and the interaction between size and AEA. The random error term assumed for the individual school is $v_{ijk} \sim N\left(0, \sigma_v^2\right)$. The model hyperparameters are $\boldsymbol{\beta}$, σ_τ^2, σ_η^2, σ_ζ^2 and σ_v^2.

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:

$$\begin{aligned} \lambda_{ijk}|\alpha, \gamma_{ijk} &\sim \text{Gamma}\left(\alpha, \alpha/\gamma_{ijk}\right) \\ \log\left(\gamma_{ijk}\right) &= \boldsymbol{\beta}'\boldsymbol{x}_{ijk} + \tau_i + \eta_j + \zeta_{ij}. \end{aligned} \tag{4.3}$$

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 is usually considered as an observed prior count when used in a prior setting. The α 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, $\boldsymbol{\beta}$, σ_τ^2, σ_η^2, and σ_ζ^2.

Under both models, the design variables AEA, district size, and school are included in the models. The sample design is considered as ignorable since it is an inherent part of the model. 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.

4.3 Hierarchical Bayes Analysis

In this section, we apply hierarchical Bayes (HB) analysis to the GLMMs introduced in Section 4.2. Estimates of the posterior mean and variance of parameters are obtained from Markov Chain Monte Carlo (MCMC) simulation.

4.3.1 Prior distributions

In a hierarchical Bayesian framework, we assume mutually independent diffuse prior distributions for the hyperparameters in (4.2) and (4.3). Let $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)$ have a (locally) uniform distribution with $p(\beta_q) \propto 1$, $q = 1, \cdots, p$. The variance component parameters are all assumed independent prior distributions: $\sigma_\tau^2 \sim \text{IG}(a_\tau, b_\tau)$, $\sigma_\eta^2 \sim \text{IG}(a_\eta, b_\eta)$, and $\sigma_\zeta^2 \sim \text{IG}(a_\zeta, b_\zeta)$, 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. The probability density function for $\text{IG}(a, b)$ is $f(x; a, b) = b^a(1/x)^a \exp(-b/x)/\Gamma(a), x > 0$, where a and b are the shape and scale parameters of the distribution. In the Poisson-Lognormal model, it is assumed that $\sigma_v^2 \sim \text{IG}(a_v, b_v)$, where a_v, and b_v are also known positive constants. The constants are usually set to be very small to reflect vague knowledge about the parameters. In our prior specification, we choose to set all the small constants to be .001. If a Poisson-Gamma model is employed, the shape parameter α can be assumed to have an independent prior distribution as $\alpha \sim \text{Gamma}(.001, .001)$. By using the proposed prior distributions, the corresponding posterior (conditional and marginal) distributions are proper.

4.3.2 MCMC sampling

The posterior distribution of unknown quantities can be approximated by replicative simulates generated using a MCMC algorithm. In the application, let $\boldsymbol{y} = (y_{ijkl})'$ denote the vector of observations. In the Poisson-Lognormal model, denote the parameters by $\boldsymbol{\theta}_{\text{p}} =$

$(\boldsymbol{\beta}', \sigma_\tau^2, \sigma_\eta^2, \sigma_\zeta^2, \sigma_v^2)'$ and the random effects by $\boldsymbol{\theta}_{\mathrm{r}} = (\boldsymbol{\tau}', \boldsymbol{\eta}', \boldsymbol{\zeta}', \boldsymbol{v}')'$. In the Poisson-Gamma model, denote the parameters by $\boldsymbol{\theta}_{\mathrm{p}} = (\alpha, \boldsymbol{\beta}', \sigma_\tau^2, \sigma_\eta^2, \sigma_\zeta^2)'$ and the random effects by $\boldsymbol{\theta}_{\mathrm{r}} = (\boldsymbol{\tau}', \boldsymbol{\eta}', \boldsymbol{\zeta}')'$. In (4.2) and (4.3), the $\boldsymbol{\lambda}$ terms are functions of the parameters and random effects. In a MCMC algorithm, the parameters and random effects are divided into blocks or components and updated componentwise by sampling from their conditional distributions given the other components. This constitutes one cycle of the update. If there are missing data or data that are unobserved by design, then as a further step in the procedure generates imputation values for the incomplete data from the model with parameters and random effects equal to their current values. Parameters and random effects then are drawn also conditional on the current values of the completed data.

If $\boldsymbol{\theta} = (\boldsymbol{\theta}_{\mathrm{p}}', \boldsymbol{\theta}_{\mathrm{r}}')'$ is partitioned into d components, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \cdots, \boldsymbol{\theta}_d')'$, then the MCMC sampling in the r^{th} iteration, $r = 1, \ldots, R$, can be implemented as follows:

(1). Generate $\boldsymbol{\theta}_1^{(r)}$ based on $f_1(\boldsymbol{\theta}_1 | \boldsymbol{\theta}_2^{(r-1)}, \cdots, \boldsymbol{\theta}_d^{(r-1)}, \boldsymbol{y}^{(r-1)})$;

(2). Generate $\boldsymbol{\theta}_2^{(r)}$ based on $f_2(\boldsymbol{\theta}_2 | \boldsymbol{\theta}_1^{(r)}, \boldsymbol{\theta}_3^{(r-1)}, \cdots, \boldsymbol{\theta}_d^{(r-1)}, \boldsymbol{y}^{(r-1)})$;

...

(d). Generate $\boldsymbol{\theta}_d^{(r)}$ based on $f_d(\boldsymbol{\theta}_d | \boldsymbol{\theta}_1^{(r)}, \cdots, \boldsymbol{\theta}_{d-1}^{(r)}, \boldsymbol{y}^{(r-1)})$;

(d+1). Generate $\boldsymbol{y}^{(r)}$ from $f(\boldsymbol{y} | \boldsymbol{\theta}^{(r)})$.

The last step in the iteration is used to generate unobserved data from the model conditional on the updated vector $\boldsymbol{\theta}$.

The set of draws are used to summarize the posterior distribution of the parameters. In the application, the MCMC sampling procedure was executed using WinBUGS and R. Simulations such as those implemented for this model are commonly implemented for hierarchical models of various specifications in other applications.

In our application, for each model the algorithm described above was implemented independently $L = 3$ times, thereby producing L parallel Markov chains. 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 $\boldsymbol{\theta}$ 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, which was sufficient to produce the Monte Carlo standard errors for all model parameters and also the deviance lower than .05.

4.3.3 Posterior estimates

Estimates of the posterior mean, variance, and covariance of $\boldsymbol{\lambda}$ 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 defined in (4.1) and (4.2) are given by

$$E(\lambda_{ijk}|\boldsymbol{y}^{obs}) = E\{\exp(\boldsymbol{\beta}'\boldsymbol{x}_{ijk} + \tau_i + \eta_j + \zeta_{ij} + \frac{1}{2}\sigma_v^2)|\boldsymbol{y}^{obs}\} \tag{4.4}$$

and

$$V(\lambda_{ijk}|\boldsymbol{y}^{obs}) = E\{\exp[2(\boldsymbol{\beta}'\boldsymbol{x}_{ijk} + \tau_i + \eta_j + \zeta_{ij} + \sigma_v^2)]|\boldsymbol{y}^{obs}\} - E^2(\lambda_{ijk}|\boldsymbol{y}^{obs}), \tag{4.5}$$

respectively. The expectations in (4.4), (4.5), (4.8) that follows are taken with respect to the joint posterior distribution of $\boldsymbol{\beta}$, $\boldsymbol{\tau}$, $\boldsymbol{\eta}$, $\boldsymbol{\zeta}$ and σ_v^2 given the data. These can be estimated using the iterated simulates from MCMC as follows:

$$\hat{E}\left(\lambda_{ijk}|\boldsymbol{y}^{obs}\right) = \frac{1}{LR}\sum_{l=1}^{L}\sum_{r=1}^{R}[\exp\{\boldsymbol{\beta}^{(lr)'}\boldsymbol{x}_{ijk} + \tau_i^{(lr)} + \eta_j^{(lr)} + \zeta_{ij}^{(lr)} + \frac{1}{2}\sigma_v^{(lr)^2}\}] \tag{4.6}$$

and

$$\hat{V}\left(\lambda_{ijk}|\boldsymbol{y}^{obs}\right) = \frac{1}{LR}\sum_{l=1}^{L}\sum_{r=1}^{R}[\exp\{2(\boldsymbol{\beta}^{(lr)'}\boldsymbol{x}_{ijk} + \tau_i^{(lr)} + \eta_j^{(lr)} + \zeta_{ij}^{(lr)} + \sigma_v^{(lr)^2})\}] - [\hat{E}(\lambda_{ijk}|\boldsymbol{y}^{obs})]^2. \tag{4.7}$$

In (4.6), (4.7), and equations 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

$$\begin{aligned} & C\left(\lambda_{ijk}, \lambda_{i'j'k'}|\boldsymbol{y}^{obs}\right) \\ = \; & C\{\exp(\boldsymbol{\beta}'\boldsymbol{x}_{ijk} + \tau_i + \eta_j + \zeta_{ij} + \frac{1}{2}\sigma_v^2), \exp(\boldsymbol{\beta}'\boldsymbol{x}_{i'j'k'} + \tau_{i'} + \eta_{j'} + \zeta_{i'j'} + \frac{1}{2}\sigma_v^2)|\boldsymbol{y}^{obs}\}. \end{aligned} \tag{4.8}$$

It can be estimated by

$$\begin{aligned} & \hat{C}\left(\lambda_{ijk}, \lambda_{i'j'k'}|\boldsymbol{y}^{obs}\right) \\ = \; & \frac{1}{LR}\sum_{l=1}^{L}\sum_{r=1}^{R}\exp\{\boldsymbol{\beta}^{(lr)'}\left(\boldsymbol{x}_{ijk} + \boldsymbol{x}_{i'j'k'}\right) + \tau_i^{(lr)} + \tau_{i'}^{(lr)} + \eta_j^{(lr)} + \eta_{j'}^{(lr)} + \zeta_{ij}^{(lr)} + \zeta_{i'j'}^{(lr)} + {\sigma_v^{(lr)}}^2\} \\ & -\hat{E}\left(\lambda_{ijk}|\boldsymbol{y}^{obs}\right)\hat{E}\left(\lambda_{i'j'k'}|\boldsymbol{y}^{obs}\right). \end{aligned} \tag{4.9}$$

If using the Poisson-Gamma model defined in (4.1) and (4.3), the posterior mean and variance of λ_{ijk} are

$$E(\lambda_{ijk}|\boldsymbol{y}^{obs}) = E\{\exp(\boldsymbol{\beta}'\boldsymbol{x}_{ijk} + \tau_i + \eta_j + \zeta_{ij})|\boldsymbol{y}^{obs}\} \tag{4.10}$$

and

$$V(\lambda_{ijk}|\boldsymbol{y}^{obs}) = E\{\exp[2(\boldsymbol{\beta}'\boldsymbol{x}_{ijk} + \tau_i + \eta_j + \zeta_{ij})(1 + 1/\alpha)]|\boldsymbol{y}^{obs}\} - E^2(\lambda_{ijk}|\boldsymbol{y}^{obs}), \tag{4.11}$$

respectively. The expectations in (4.10), (4.11), and (4.14) that follows are taken with respect to the joint posterior distribution of $\boldsymbol{\beta}$, $\boldsymbol{\tau}$, $\boldsymbol{\eta}$, $\boldsymbol{\zeta}$ and α given the data. They can be estimated using the iterated simulates from MCMC as follows:

$$\hat{E}\left(\lambda_{ijk}|\boldsymbol{y}^{obs}\right) = \frac{1}{LR}\sum_{l=1}^{L}\sum_{r=1}^{R}[\exp\{\boldsymbol{\beta}^{(lr)'}\boldsymbol{x}_{ijk} + \tau_i^{(lr)} + \eta_j^{(lr)} + \zeta_{ij}^{(lr)}\}] \tag{4.12}$$

and

$$\hat{V}\left(\lambda_{ijk}|\boldsymbol{y}^{obs}\right) = \frac{1}{LR}\sum_{l=1}^{L}\sum_{r=1}^{R}[\exp\{2(\boldsymbol{\beta}^{(lr)'}\boldsymbol{x}_{ijk} + \tau_i^{(lr)} + \eta_j^{(lr)} + \zeta_{ij}^{(lr)})(1 + 1/\alpha^{(lr)})\}] - [\hat{E}(\lambda_{ijk}|\boldsymbol{y}^{obs})]^2. \tag{4.13}$$

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

$$C\left(\lambda_{ijk}, \lambda_{i'j'k'}|\boldsymbol{y}^{obs}\right) = C\{\exp(\boldsymbol{\beta}'\boldsymbol{x}_{ijk} + \tau_i + \eta_j + \zeta_{ij}, \exp(\boldsymbol{\beta}'\boldsymbol{x}_{i'j'k'} + \tau_{i'} + \eta_{j'} + \zeta_{i'j'})|\boldsymbol{y}^{obs}\}. \tag{4.14}$$

It can be estimated by

$$\begin{aligned}
&\hat{C}\left(\lambda_{ijk}, \lambda_{i'j'k'}|\boldsymbol{y}^{obs}\right) \\
= \;& \frac{1}{LR}\sum_{l=1}^{L}\sum_{r=1}^{R}\exp\{\boldsymbol{\beta}^{(lr)\prime}\left(\boldsymbol{x}_{ijk}+\boldsymbol{x}_{i'j'k'}\right)' + \tau_i^{(lr)} + \tau_{i'}^{(lr)} + \eta_j^{(lr)} + \eta_{j'}^{(lr)} + \zeta_{ij}^{(lr)} + \zeta_{i'j'}^{(lr)}\} \\
& -\hat{E}\left(\lambda_{ijk}|\boldsymbol{y}^{obs}\right)\hat{E}\left(\lambda_{i'j'k'}|\boldsymbol{y}^{obs}\right). \qquad (4.15)
\end{aligned}$$

The derivation for (4.4), (4.5), (4.8), (4.10), (4.11), and (4.14) is given in an Appendix.

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 (i,j) is $N_{ij} = \sum_{k\in U_{ij}} N_{ijk}$, where N_{ijk} is the number of students in the k^{th} school in the stratum.

The average μ_{ij} can be considered as the sum of three terms:

$$\mu_{ij} \;=\; N_{ij}^{-1}\{\sum_{k\in s_{ij}}\sum_{l\in s_{ijk}}\tilde{Y}_{ijkl} + \sum_{k\in s_{ij}}\sum_{l\notin s_{ijk}}\tilde{Y}_{ijkl} + \sum_{k\notin s_{ij}}\sum_{l\in U_{ijk}}\tilde{Y}_{ijkl}\}, \qquad (4.16)$$

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

$$\begin{aligned}
& E\left(\mu_{ij}|\boldsymbol{y}^{obs}\right) \\
= \;& N_{ij}^{-1}\{\sum_{k\in s_{ij}} 8\sum_{l\in s_{ijk}} y_{ijkl}/\omega_{ijkl} + \sum_{k\in s_{ij}} 8(N_{ijk}-n_{ijk})E\left(\lambda_{ijk}|\boldsymbol{y}^{obs}\right) + \sum_{k\notin s_{ij}} 8N_{ijk}E\left(\lambda_{ijk}|\boldsymbol{y}^{obs}\right)\} \\
\equiv \;& N_{ij}^{-1}\{\sum_{k\in s_{ij}} 8\sum_{l\in s_{ijk}} y_{ijkl}/\omega_{ijkl} + \boldsymbol{l}_{ij}'E\left(\boldsymbol{\lambda}|\boldsymbol{y}^{obs}\right)\}.
\end{aligned}$$

In the above, $\boldsymbol{\lambda} = \{\lambda_{ijk}\}$ is a parameter vector of Poisson distribution rates for schools in the entire population and $\boldsymbol{l}_{ij} = \{\boldsymbol{0}', \cdots, \boldsymbol{0}', \tilde{\boldsymbol{l}}_{ij}', \boldsymbol{0}', \cdots, \boldsymbol{0}'\}'$ is the vector of coefficients for stratum (i,j). In the latter expression, $\tilde{\boldsymbol{l}}_{ij} = \{l_{ijk}\}_{k\in U_{ij}}$ is the vector of values l_{ijk} in stratum (i,j). The value l_{ijk} equals $8(N_{ijk}-n_{ijk})$ if $k \in s_{ij}$, and equals $8N_{ijk}$ if $k \notin s_{ij}$.

The proposed HB estimator of μ_{ij} is

$$\hat{\mu}_{ij} = N_{ij}^{-1}\{\sum_{k\in s_{ij}} 8 \sum_{l\in s_{ijk}} y_{ijkl}/\omega_{ijkl} + \boldsymbol{l}'_{ij}\hat{E}\left(\boldsymbol{\lambda}|\boldsymbol{y}^{obs}\right)\}. \tag{4.17}$$

The posterior variance of $\hat{\mu}_{ij}$ is

$$V\left(\mu_{ij}|\boldsymbol{y}^{obs}\right) = N_{ij}^{-2}\{\boldsymbol{l}'_{ij}V\left(\boldsymbol{\lambda}|\boldsymbol{y}^{obs}\right)\boldsymbol{l}_{ij}\}, \tag{4.18}$$

which can be estimated by plugging $\hat{V}\left(\boldsymbol{\lambda}|\boldsymbol{y}^{obs}\right)$ into (4.18). The diagonal and off-diagonal elements of $\hat{V}\left(\boldsymbol{\lambda}|\boldsymbol{y}^{obs}\right)$ are calculated by (4.7) and (4.9) under the Poisson-Lognormal model and by (4.13) and (4.15) under the Poisson-Gamma model, respectively.

4.4 Illustration

To illustrate the performance of the hierarchical Bayesian estimation method, we simulated a finite population of EP courses taken by twelfth grade students from Iowa's public high schools from a Poisson log-linear model:

$$\begin{aligned} y_{ijkl}|\lambda_{ijk} &\sim \text{Poisson } (\lambda_{ijk}) \\ \log(\lambda_{ijk}) &= \beta_0 + \beta_1 x_{ijk;1} + \tau_i + \eta_j. \end{aligned}$$

The number of EP courses taken by the $l-$th student in the $k-$th school in stratum (i,j) follows a Poisson distribution with mean λ_{ijk}, where λ_{ijk} denotes the underlying school-specific rate of taking EP courses. 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 log-rate of taking EP courses is then assumed to be linearly related to some covariable variable at the school and the random effects from size levels and AEAs. The logarithm of the enrollment size in twelfth grade was used as an auxiliary variable $\boldsymbol{x}_1$ in generating the population data set. An intercept term β_0 was included in the log-linear model. The τ_i, $i = 1, \cdots, I = 3$ and η_j, $j = 1, \cdots, J = 12$ are the size and AEA random effects. Population sizes in the simulation match actual population sizes in Iowa's school districts in 2004. The values of the parameters that were used in generating the population are as follows: $\beta_0 = .5$,

$\beta_1 = .2$, $\boldsymbol{\tau} = (.5, .2, -.1)$, $\boldsymbol{\eta} = (.3, .25, .2, .15, .1, .05, -.05, -1, . - 15, -.2, -.25, -.3)$, $\sigma_\tau = .3$, $\sigma_\eta = .25$. The values of $\boldsymbol{x}_1$ ranges from 3 to 6 with the mean around 4.75. One sample data set was drawn from the simulated population using the stratified three-stage design described in Section 2.1.

Both direct and hierarchical Bayesian estimation were used to produce the estimates of the stratum means. In the hierarchical Bayesian analysis, the prior distributions for model parameters were specified as follows: β_0 follows a normal distribution with mean 0 and a large variance (10^4), independently β_1 has a uniform distribution, and $\sigma_\tau^2, \sigma_\eta^2 \sim IG(.001, .001)$. The marginal and conditional posterior distributions are proper. 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 were deleted as a "burn-in" period. By thinning to every 5^{th} iteration, 1,000 iterates from each chain were retained for posterior estimation. The HB estimates and the posterior variances were calculated using the retained simulates from the approximated posterior distribution by MCMC simulation.

The comparison between the model-based HB estimator with the design-based ratio estimator were based on the absolute relative bias (ARB) and the root of mean squared 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 the posterior variance under the assumed model.

Figure 4.1 shows the absolute relative bias (ARB) of ratio and HB estimators over the realized (true) finite population mean for strata of medium districts. 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 absolute bias 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

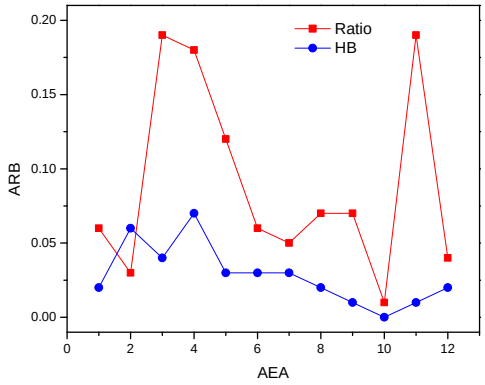


Figure 4.1 The Absolute Relative Bias (ARB) of ratio and HB estimators for strata of medium districts based on a sample data set from a single simulated finite population.

sampled. Also the ratio estimator shows much higher variation than the HB estimator at the small area (stratum) level. The results for small strata were qualitatively similar. The ratio estimator had much larger ARBs in most of the strata and showed larger variation overall.

Figure 4.2 displays the root of mean squared error (RMSE) of the ratio estimator and the root of posterior mean squared error (RPMSE) of the HB estimator for the medium strata. The MSE of the ratio estimator was estimated through 1,000 replicated simulations of the sample from the finite population. The posterior MSE of the HB estimator was derived under the true model that was used to generate the finite population based on the sample data set that was used to evaluate the ARBs of the estimators. The RMSE of the ratio estimator is consistently higher than the RPMSE of the HB estimator for all but one stratum of medium districts. The RPMSE of the HB estimator was no more than .2 in all medium strata. However, the RMSE of the ratio estimator is at least twice higher than the RPMSE of the HB estimator in most of the strata. Therefore, the ratio estimator produced much more variation in estimating the medium stratum means than the HB estimator. However, the advantage of the HB estimator in terms of producing consistent and significant higher precision was not observed in strata of small districts. The reason is that the enrollment size of twelfth grade students in small school districts has very small range, hence the rates of taking EP courses in small districts within

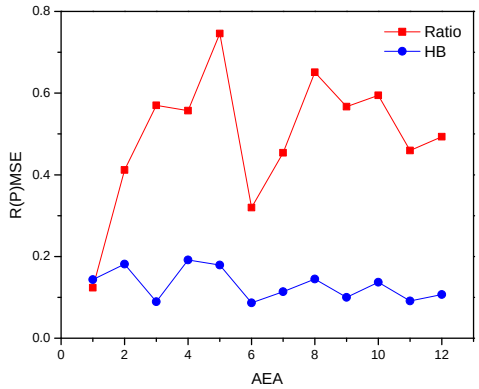


Figure 4.2 The root of mean squared error (RMSE) of the ratio estimator and the root of posterior mean squared error (RPMSE) of the HB estimator for strata of medium districts. The RMSE of ratio estimate was obtained based on 1,000 simulated samples.

each stratum are very close. Since there is weak homogeneity within the PSUs in each small stratum, the direct estimator would not produce very imprecise estimate even based on a very small size sample.

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 4.3 shows two kinds of standard errors (SEs) of the ratio estimator for strata with one PSU sampled. The RA1 was obtained by collapsing strata followed by synthetic variance redistribution, and is the CSSV method of Chapter 3. The RA2 was estimated by using the collapsing strata restricted generalized variance function synthetic variance estimation, and is the CRGVFSV method of Chapter 3. In the case of our application, the collapsing strata estimator significantly overestimated the variances in small areas with one-PSU sampled. 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

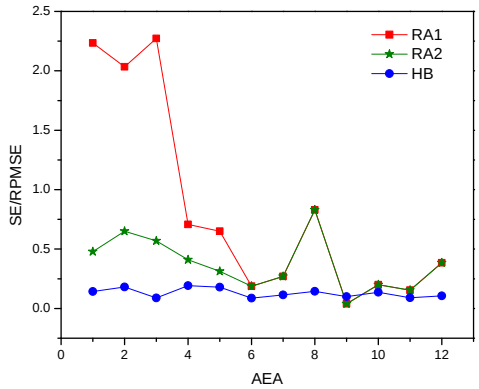


Figure 4.3 The standard errors (SEs) of the ratio estimate and RPMSE of the HB estimate for strata of medium districts based on a sample data set from a simulated finite population. The SEs of ratio estimate was obtained by using collapsed strata synthetic variance (CSSV) and collapsed strata generalized variance function synthetic variance (CRGVFSV) methods, denoted by RA1 and RA2 respectively.

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 RPMSE. The advantage of using the HB estimator is significant in terms of producing more accurate and reliable estimate and a better assessment of precision of the estimate. 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.

4.5 ISBE Survey Data Analysis

The ISBE survey was designed as a stratified multi-stage survey. The stratification factors are the size of school district and AEA. There are three size levels and 12 AEAs, hence the whole state is technically stratified into $3 \times 12 = 36$ strata. However, not every defined

stratum actually contains population units. There are 22 schools from eight large districts in seven AEAs, 144 schools from 143 medium districts and 177 schools from 177 small districts in twelve AEAs. Five AEAs do not have large school districts. So the population is actually divided into 31 strata among which seven have large school districts, twelve have medium and twelve have small districts. A large district has at least two high schools. The largest one has five schools in supervision. All small districts and all except one medium school district has only one school each.

The survey was planned to sample 60 schools and no more than 12,000 students from the whole state. The 22 schools from large districts were about to be taken with certainty. In each of the medium and small size levels, 19 schools were supposed to be taken from 12 AEAs, among which 14 schools were from 7 AEAs that have more districts and 5 schools were from the remaining 5 AEAs with fewer districts. The school districts in medium and small strata were sampled by proportional to size sampling without replacement based on the total enrollment of grade nine to grade twelve students. Students within each sampled school would be selected by sampling random sampling without replacement. Approximately 50 students on average were about to be selected from each of the four groups from grade nine or grade twelve and having general or special education within each sample school.

The actual survey data were collected in 2005. The data contain over 4,000 students from 51 sample schools in 11 AEAs. AEA 9 and large and small school districts in AEA 1 did not participate. So there were only 26 strata with sample schools in the actual data: five strata of large school districts, 11 strata with medium school districts and 10 strata containing small districts. Among school districts that participate in the survey, all large districts and one medium district in AEA 1 had more than one schools. The rest of the districts (medium or small) have only one school each. There are 17 sample schools in each of the size levels in the actual survey data. In strata with medium districts, six strata had a single school, four strata had two schools each and one stratum (which has the most districts) had three schools in the sample. In strata with small districts, three strata had only one school sampled and seven strata had two sample schools each. The sample and population sizes of school districts within

each stratum are summarized in Table 4.1.

Table 4.1 The numbers of schools districts within each stratum in the actual sample data and in the population. "M" means there are school districts sampled in the stratum which refused to participate in the survey. "None" means no school districts existing in the stratum.

Sample/Population	Large	Medium	Small
AEA 1	M/1	2/12	M/9
AEA 4	None	1/6	1/7
AEA 8	None	2/16	2/27
AEA 9	M/1	M/13	M/7
AEA 10	2/2	2/15	2/14
AEA 11	1/1	3/29	2/22
AEA 12	1/1	1/6	2/13
AEA 13	1/1	1/7	2/21
AEA 14	None	1/4	2/13
AEA 15	None	1/6	1/5
AEA 16	None	1/6	1/5
AEA 267	1/1	2/23	2/34

Both design-based estimation and hierarchical Bayesian (HB) estimation were applied to analyze the survey data. The Horvitz-Thompson (HT) and ratio estimators were used in design-based estimation. The variances of estimators for strata with at least two PSUs sampled were estimated using direction variance estimation for a stratified multi-stage design. The variances of estimators for one-PSU strata are estimated using collapsed strata variance estimator followed by synthetic variance redistribution (CSSV) and collapsed strata restricted generalized variance function synthetic variance (CRGVFSV) estimation methods. The estimators and variance estimation methods showed similar properties as they did in the previous simulation study (Lu and Larsen (2006, 2007)). More details about the results of the previous simulation study using design-based estimation were shown in Chapters 2 and 3. In this section, we are going to conduct hierarchical Bayesian analysis of the ISBE survey data using some GLMMs described in Section 4.2. The posterior estimates of the model parameters and of the stratum means (the small area quantities of interest) are summarized. The results obtained

using several GLMMs with different model assumptions are compared. Further discussion of a formal model selection among the candidate models will be discussed in Chapter 5.

An informal model building process was done in hierarchical Bayesian analysis to the ISBE survey data. Under the two families of models that were being considered, a variety of models that assume different nature and source of variabilities and number of auxiliary variables were examined. The models were built in an iterative procedure to choose promising models from a class of plausible models. Some initial models were chosen and fitted to the data. Then the models were checked for possible deficiency, and the next candidate models were proposed based on the model checking results of the current model. The process was continued until a model or a few models were identified as the most satisfactory models to describe the survey data and make inferences and predictions in the subject matter.

For the Poisson-Gamma models, we started with the simplest model which assumes no variability among the school-specific log-rate of taking EP classes and no relationship to any administrative variables. The model no surprisingly produced very poor estimates of stratum means and extremely large variance estimates. The HB estimates for all strata were bigger than 6 and most of them were bigger than 7, which was quite impossible because the reliable design-based state mean was only 5.4 and the HB estimates based on the simplest model were systematically dragged to the larger values. So the simplest (pooled) model was shown to have too large lack-of-fit and we would like to consider more complex models.

Based on the design-based estimates of the size levels, we observed an increasing pattern in the average number of EP courses taken as the size of the school districts is getting smaller. This suggested to include the factor of the size of school districts into the model of the log-school rates. There are two ways of bringing in the size factor. One way is by assuming a linear relationship with the school enrollment size variable which will only add one more parameter into the model. The other is by including a size random effect term which will add three more parameters. The HB estimates under the model with the auxiliary variable were closely gathered around the state mean and did not show enough variation among the size levels, which suggested the model is over-shrunk and the "size" factor should be included in a

stronger way.

So a model with size random effect was then fitted to the data. The HB estimates show clear distinctions among size levels and the estimates for the small size level seemed to be systematically lower than the design-based estimate of the small size level. Also, for some of the AEAs such as AEAs 8 and 10, there were two PSUs sampled in each stratum of different size levels. The direct estimates of these AEAs are design-consistent and less unreliable compared to the estimates for other AEAs that contain one-PSU-per-stratum strata. The direct estimate for AEA 8 was 9.29 which was much bigger than the HB estimate 6.24, and the direct estimate for AEA 10 was 2.19 which was only one half of the HB estimate 4.39. This provided "evidence" that the model with only size random effect failed to capture some prominent variation among AEAs. Hence, a more complex model that includes both size and AEA random effects was suggested for further examination.

Table 4.2 shows the posterior estimates (including the posterior mean, the standard error, the median, the 95% credible interval and the Monte Carlo standard error) of parameters in the Poisson-Gamma model with size and AEA random effects and no auxiliary variable. The model captured the significant trend that larger school districts tend to have students who on average take fewer EP courses. The 95% credible intervals of τ_{large} was significantly away from zero. The credible interval of τ_{medium} only marginally covered zero. The credible intervals for random effects for AEAs 10 and 16 were also significantly away from zero and marginally covered zero for AEA 8. The directions of the deviations of the credible intervals from zero exactly match the pattern of deviations of direct estimates of the means of the certain levels from the state mean.

Another way of including both size and AEA factors in the model of log-school rates is through a model that has a random coefficient for size levels and AEA random effects. Compared to the model with both size and AEA random effects, the random coefficient model does not add to the number of parameters, but allows variation in regression coefficients across size levels and thus avoids the "over-shrinkage" along the size factor dimension. Table 4.3 shows the posterior estimates of parameters in the Poisson-Gamma model with a random

coefficient for size levels and AEA random effects. Like the model with both size and AEA random effects, the random coefficient model also captured the pattern that students from larger school districts tend to take fewer EP courses. More specifically, students from large school districts tend to take fewer EP courses than students from medium and small schools. Also within each size level, students from districts with larger enrollment size have a tendency to take fewer EP courses. The posterior credible intervals for random coefficients $\beta_{1,\text{large}}$ and $\beta_{1,\text{medium}}$ for large and medium size levels did not include 0. The credible interval for $\beta_{1,\text{small}}$ only included 0 near the boundary. The credible intervals for random effects for AEAs 8, 10 and 16 also did not cover 0. The interval for AEA 14 included 0 near the boundary. The signs of posterior medians were also consistent with the signs of direct estimates of the AEA means after subtracting the state mean.

The model with both size and AEA random effects and the model with a random coefficient for size levels and AEA random effects both captured the main features of the observed data. The posterior estimates of the model parameters convey consistent information about the population characteristics. Table 4.4 shows the HB estimates of stratum means and the coefficient of variations (CVs) of the HB estimates for the Poisson-Gamma model with size and AEA random effects and no auxiliary variable (Model 1) and the Poisson-Gamma model with a random coefficient for size levels and AEA random effects (Model 2). The HB estimates derived from the two models were very close to one another. The CVs of the HB estimates from the random coefficient model were slightly bigger than the CVs from model with only random effects for all small areas and most of the medium strata. However, the difference is generally only to the third decimal place which is very small. Therefore, the models were indistinguishable in terms of HB estimates of stratum means.

Table 4.4 also shows the ratio estimates and the CVs of ratio estimates which were calculated from the variance estimates obtained using the collapsed strata synthetic variance (CSSV) and the collapsed strata generalized variance function synthetic variance (CRGVFSV) methods, which are denoted by CV_1 and CV_2, respectively. Compared to the CVs of the ratio estimates using the generalized variance function method, the CVs of HB estimates were con-

sistently smaller than the CVs of ratio estimates for all but one medium strata and more than half of the small strata. In particular, the CVs were significantly reduced using HB estimator for all but one strata with a single PSU sampled. The CSSV method generally produced larger variance estimates than the generalized variance function method for most of the one-PSU strata. Therefore, the reduction in the CVs were more prominent for most of strata with a single PSU in the sample. This suggests that the HB estimator under the two models produces more precise and reliable estimates of stratum means.

The model building procedure for the Poisson-Lognormal models was quite similar. We started with the simplest (pooled) model which fitted the data very poorly and increased the complexity of the model one step ahead. The expanded model was fitted and examined by checking the posterior estimates and identifying some inadequacy of the model. The process then involved proposing a new model incorporating more administrative variables or sources of variability. The MCMC simulation for the Poisson-Lognormal models generally took a longer time than for the Poisson-Gamma models. The mixing process was much slower and the iterative draws were highly correlated, hence the simulation required much longer Markov chains and more time for running the process. Therefore, for each model, we simulated 200,000 iterations. After the first 100,000 iterations of "burn-in" period, 1,000 values thinned by every 100th iteration were retained for posterior estimation. The Monte Carlo standard error for all model parameters were less than 0.02.

After this iterative procedure, we ended up with the model that has both size and AEA random effects. Table 4.5 shows the posterior estimates of parameters in the Poisson-Lognormal model with size and AEA random effects. The 95% posterior credible interval for τ_{large} did not cover zero and was to the negative side of zero. The credible interval for τ_{medium} did include 0, which was on the boundary of the interval. We can see clearly the increasing pattern of the posterior means and medians of the size random effect τ_i as the size level goes from large to small. The credible intervals of η_{10} and η_{16} for AEAs 10 and 16 did not include 0. The interval for η_8 covered 0 on the boundary. The signs of the posterior means of η_8, η_{10} and η_{16} also matched the directions of the deviation of the design-based estimates of these AEAs from

the state mean. All these patterns were consistent with the patterns we observed under the Poisson-Gamma model with size and AEA random effects. We also compared the posterior estimates of stratum means obtained under the two families of models (see Table 4.6). The HB estimates of means did not differ very much for most strata. The coefficient of variations were also very comparable. As a summary, the Poisson-Lognormal model with size and AEA random effects produced the posterior estimates of stratum means almost as precise as the Poisson-Gamma model with both size and AEA random effects. However, the mixing in the MCMC simulation for the Poisson-Lognormal model was much slower and the autocorrelation of iterative draws of the random effect parameters were much higher than for the Poisson-Gamma model. Hence it required running much longer Markov chains to approximate the posterior distribution and the simulation is much more time-consuming. Therefore, we prefer to use the Poisson-Gamma model with size and AEA random effects or the Poisson-Gamma model with random coefficients for size levels and AEA random effects in further data analysis.

4.6 Summary and Discussion

The data analysis in the ISBE survey raised a problem of improving the precision of the design-based estimator due to the smallness of the sample size within each stratum in the survey. Also for extreme cases like the ISBE survey in which a single PSU is sampled within each stratum, variance estimation is another big challenge. It is desirable to use a better estimation method that can produce not only more precise and reliable direct estimates but also more reliable assessments of the precision of the estimates. An examination of small area estimation through model-based inference was motivated by the survey. The method of producing more reliable estimates for areas with small sample sizes than direct estimation were studied from a full Bayesian perspective.

Two families of generalized linear mixed models (GLMMs), Poisson-Gamma and Poisson-Lognormal models, were proposed for modeling the ISBE survey data. 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.

In the illustrative example, a finite population of EP courses taken by twelfth grade students from Iowa's public high schools was simulated. A single sample was drawn from the simulated population under the stratified multi-stage design. The design-based ratio estimator and the hierarchical Bayesian estimator were applied to produce the estimates of the stratum means for the sample data. The HB estimator outperformed the ratio estimator by "borrowing strength" across related strata in terms of producing consistently smaller absolute relative bias (ARB) to the realized finite population mean and root of (posterior) mean square error (R(P)MSE) for individual strata.

The model building for analyzing the actual survey data was conducted in an informal iterative procedure within each family of GLMMs. A simple model was initiated and fitted to the data, and then checked for model inadequacy and a new model was proposed if any inadequacy was detected. The complexity of the examined model was increased along the cycle of model fitting, checking and updating until a single model or a couple of most satisfactory models were selected. The analysis showed strong evidence of variability along both size and AEA factors. The Poisson-Gamma model with size and AEA random effects and the Poisson-Gamma model that has enrollment size as an auxiliary variable with a random coefficient for size levels and AEA random effects were considered the best models within the family. The Poisson-Lognormal model with size and AEA random effects was chosen from its family. The posterior estimates under the Poison-Gamma and the Poisson-Lognormal models are very consistent and comparable. However, considering that the MCMC simulation is much more time-consuming for the Poisson-Lognormal model and the Poisson-Gamma model is in fact more robust and provides minimax estimators for mean square error (Christiansen and Morris (1997)), we would prefer the Poisson-Gamma models.

The hierarchical analysis based on the Poisson-Gamma model with size and AEA random effects showed that there is a tendency that students from large school districts will take fewer EP courses than students from medium and small school districts. Also within each size level, students from larger school districts tend to select fewer EP courses than students from

smaller districts. The variability among AEAs is also a very important factor to be considered in analysis of these data.

Table 4.2 Posterior estimates of parameters in the Poisson-Gamma model with size and AEA random effects and no auxiliary variable: $\log(\gamma_{i,j,k}) = \beta_0 + \tau_i + \eta_j$. The bold figures are the 95% credible interval bounds for the intervals that exclude 0. The $\tau_i, i \in \{\text{large, medium, small}\}$ denote size random effects. The $\eta_j, j = 1, 4, 8, 9, 10, 11, 12, 13, 14, 15, 16, 267$ represent AEA random effects. The subscript index the actual size levels and AEAs.

	Mean	Standard Error	2.5% Percentile	Median	97.5% Percentile
β_0	-.0005	.0099	-.0208	-.0006	.0187
$\boldsymbol{\tau}_{\text{large}}$	-.7325	.1880	**-1.0580**	-.7506	**-.3209**
τ_{medium}	-.3394	.1695	-.6435	-.3549	.0269
τ_{small}	-.2929	.1734	-.5991	-.3074	.0885
η_1	.1092	.2249	-.3596	.1167	.5188
η_4	-.1878	.2328	-.6809	-.1813	.2496
η_8	.3856	.1938	-.0203	.3925	.7571
η_9	.0038	.4601	-.9000	.0064	.9573
$\boldsymbol{\eta}_{10}$	-.6864	.2011	**-1.1420**	-.6727	**-.3239**
η_{11}	-.0183	.1836	-.4185	-.0060	.3149
η_{12}	-.0082	.1938	-.4304	.0060	.3406
η_{13}	.2698	.1930	-.1396	.2807	.6260
η_{14}	-.3349	.2224	-.8138	-.3204	.0749
η_{15}	-.1209	.2329	-.6033	-.1110	.3221
$\boldsymbol{\eta}_{16}$	-.7095	.2745	**-1.2810**	-.6938	**-.1897**
η_{267}	.1228	.1889	-.2751	.1338	.4728
σ_τ	.7103	.6402	.1888	.5599	2.0430
σ_η	.4350	.1408	.2386	.4111	.7964
α	13.5500	3.2270	8.1610	13.2300	20.7300

Table 4.3 Posterior estimates of parameters in the Poisson-Gamma model with AEA random effect and an auxiliary variable with random coefficient for size levels: $\log(\gamma_{i,j,k}) = \beta_0 + \beta_{1,i}x_{i,j,k} + \eta_j$.

	Mean	Standard Error	2.5% Percentile	Median	97.5% Percentile
β_0	-.0006	.0101	-.0204	-.0006	.0189
$\boldsymbol{\beta}_{1,\text{large}}$	-.1261	.0301	**-.1804**	-.1280	**-.0614**
$\boldsymbol{\beta}_{1,\text{medium}}$	-.0681	.0314	**-.1255**	-.0692	**-.0043**
$\beta_{1,\text{small}}$	-.0733	.0432	-.1531	-.0755	.0135
η_1	.1021	.2179	-.3344	.1050	.5183
η_4	-.1941	.2262	-.6522	-.1910	.2371
$\boldsymbol{\eta}_8$	.3715	.1879	**.0072**	.3718	**.7339**
η_9	.0005	.4554	-.9019	-.0027	.9180
$\boldsymbol{\eta}_{10}$	-.6793	.1900	**-1.0710**	-.6714	**-.3303**
η_{11}	-.0223	.1735	-.3759	-.0127	.2967
η_{12}	-.0143	.1827	-.3793	-.0067	.3407
η_{13}	.2403	.1768	-.1179	.2445	.5812
η_{14}	-.3624	.2123	-.7772	-.3603	.0353
η_{15}	-.1479	.2216	-.5949	-.1413	.2741
$\boldsymbol{\eta}_{16}$	-.7050	.2753	**-1.2590**	-.6997	**-.1821**
η_{267}	.1019	.1711	-.2422	.1105	.4157
σ_β	.1383	.1221	.0395	.1103	.4009
σ_η	.4301	.1351	.2350	.4064	.7649
α	13.3800	3.1740	7.9180	13.1000	20.3300

Table 4.4 The HB estimates of stratum means and the coefficient of variation (CV) of the HB estimates for Poisson-Gamma model with size and AEA random effects and no auxiliary variable (Model 1): $\log(\gamma_{i,j,k}) = \beta_0 + \tau_i + \eta_j$ and Poisson-Gamma model with a random coefficient for size levels and AEA random effect (Model 2): $\log(\gamma_{i,j,k}) = \beta_0 + \beta_{1,i}x_{i,j,k} + \eta_j$. The CVs for the ratio estimator are calculated from the variance estimates obtained using the collapsed strata synthetic variance (CSSV) and the collapsed strata generalized variance function synthetic variance (CRGVFSV) methods, which are denoted by CV_1 and CV_2 respectively.

	Model 1		Model 2		Ratio Estimator		
AEA	$\hat{\mu}_{ij}^{HB}$	$CV(\hat{\mu}_{ij}^{HB})$	$\hat{\mu}_{ij}^{HB}$	$CV(\hat{\mu}_{ij}^{HB})$	$\hat{\mu}_{ij}^{RA}$	$CV_1(\hat{\mu}_{ij}^{RA})$	$CV_2(\hat{\mu}_{ij}^{RA})$
			Strata of medium districts				
1	6.42	0.1704	6.49	0.1746	6.46	0.1708	0.1708
4	4.61	0.1986	4.73	0.1990	3.77	0.1564	0.3447
8	8.56	0.1458	8.54	0.1515	9.30	0.2384	0.2384
10	2.77	0.1421	2.80	0.1424	2.10	0.2430	0.2430
11	5.58	0.1200	5.52	0.1220	4.96	0.0173	0.0173
12	5.87	0.1443	5.95	0.1432	6.50	0.6899	0.4616
13	8.23	0.1427	8.12	0.1430	12.16	0.5252	0.3514
14	4.09	0.1909	4.09	0.1949	3.91	0.2755	0.6070
15	5.22	0.2080	5.03	0.2110	5.73	0.5339	0.2743
16	2.79	0.2464	2.78	0.2539	2.23	0.8173	0.4198
267	6.49	0.1385	6.40	0.1374	6.92	0.1692	0.1692
			Strata of small districts				
4	5.11	0.1884	5.12	0.1890	5.44	0.6219	0.2889
8	8.91	0.1474	9.05	0.1519	9.30	0.0849	0.0849
10	2.99	0.1396	3.06	0.1448	2.61	0.1124	0.1124
11	6.02	0.1163	6.11	0.1182	7.12	0.3622	0.3622
12	6.07	0.1301	6.09	0.1311	6.64	0.0085	0.0085
13	7.70	0.1428	7.63	0.1460	6.57	0.2728	0.2728
14	4.31	0.1622	4.31	0.1692	3.92	0.1617	0.1617
15	5.26	0.2311	5.26	0.2366	4.20	0.4384	0.2036
16	2.90	0.2256	2.96	0.2348	2.46	0.8280	0.3846
267	6.80	0.1369	6.75	0.1381	6.49	0.2988	0.2988

Table 4.5 Posterior estimates of parameters in Poisson-Lognormal model with size and AEA random effects and no auxiliary variable: $\log(\lambda_{i,j,k}) = \beta_0 + \tau_i + \eta_j + v_{i,j,k}$.

	Mean	Standard Error	2.5% Percentile	Median	97.5% Percentile
β_0	-.0007	.0099	-.0203	-.0006	.0184
$\boldsymbol{\tau}_{\text{large}}$	-.7333	.2044	**-1.0850**	-.7564	**-.2781**
τ_{medium}	-.3595	.1820	-.6773	-.3798	.0387
τ_{small}	-.3077	.1857	-.6311	-.3252	.0920
η_1	.1060	.2313	-.3703	.1177	.5415
η_4	-.2075	.2452	-.7315	-.1967	.2419
η_8	.3687	.2054	-.0592	.3780	.7578
η_9	.0124	.4745	-.9296	.0076	1.0020
$\boldsymbol{\eta}_{10}$	-.7280	.2096	**-1.1800**	-.7095	**-.3660**
η_{11}	-.0419	.1987	-.4795	-.0187	.3126
η_{12}	-.0165	.2065	-.4594	-.0036	.3646
η_{13}	.2233	.2043	-.2104	.2329	.6021
η_{14}	-.3491	.2315	-.8262	-.3375	.0694
η_{15}	-.1340	.2394	-.6364	-.1198	.3197
$\boldsymbol{\eta}_{16}$	-.7349	.2713	**-1.2940**	-.7188	**-.2423**
η_{267}	.1062	.2052	-.3329	.1191	.4755
σ_τ	.7059	.5596	.1732	.5713	2.1660
σ_η	.4516	.1525	.2434	.4197	.8688
σ_v	.2758	.0346	.2176	.2728	.3535

Table 4.6 Posterior estimates of stratum means for Poisson-Lognormal model with size and AEA random effects and no auxiliary variable: $\log(\lambda_{i,j,k}) = \beta_0 + \tau_i + \eta_j + v_{i,j,k}$.

	Poisson-Gamma		Poisson-Lognormal		Ratio Estimator		
AEA	$\hat{\mu}_{ij}^{HB}$	$\text{CV}(\hat{\mu}_{ij}^{HB})$	$\hat{\mu}_{ij}^{BHB}$	$\text{CV}(\hat{\mu}_{ij}^{BHB})$	$\hat{\mu}_{ij}^{RA}$	$\text{CV}_1(\hat{\mu}_{ij}^{RA})$	$\text{CV}_2(\hat{\mu}_{ij}^{RA})$
			Strata of medium districts				
1	6.42	0.1704	6.50	0.1759	6.46	0.1708	0.1708
4	4.61	0.1986	4.60	0.1953	3.77	0.1564	0.3447
8	8.56	0.1458	8.57	0.1480	9.30	0.2384	0.2384
10	2.77	0.1421	2.71	0.1354	2.10	0.2430	0.2430
11	5.58	0.1200	5.55	0.1200	4.96	0.0173	0.0173
12	5.87	0.1443	5.91	0.1446	6.50	0.6899	0.4616
13	8.23	0.1427	8.06	0.1467	12.16	0.5252	0.3514
14	4.09	0.1909	4.10	0.1918	3.91	0.2755	0.6070
15	5.22	0.2080	5.25	0.2140	5.73	0.5339	0.2743
16	2.79	0.2464	2.76	0.2347	2.23	0.8173	0.4198
267	6.49	0.1385	6.50	0.1381	6.92	0.1692	0.1692
			Strata of small districts				
4	5.11	0.1884	5.12	0.1890	5.44	0.6219	0.2889
8	8.91	0.1474	8.97	0.1493	9.30	0.0849	0.0849
10	2.99	0.1396	2.94	0.1379	2.61	0.1124	0.1124
11	6.02	0.1163	6.03	0.1205	7.12	0.3622	0.3622
12	6.07	0.1301	6.14	0.1295	6.64	0.0085	0.0085
13	7.70	0.1428	7.54	0.1425	6.57	0.2728	0.2728
14	4.31	0.1622	4.34	0.1638	3.92	0.1617	0.1617
15	5.26	0.2311	5.31	0.2344	4.20	0.4384	0.2036
16	2.90	0.2256	2.88	0.2150	2.46	0.8280	0.3846
267	6.80	0.1369	6.84	0.1316	6.49	0.2988	0.2988

CHAPTER 5. Hierarchical Bayesian Model Selection Using Benchmarking

In Chapter 4, we used hierarchical Bayesian methods to produce reliable estimates of small area quantities (stratum means). Different generalized linear mixed models were considered for capturing the hierarchical structure in the stratified multi-stage sample survey. The models with different random effects and forms of involving covariable variables were fitted and their performances were evaluated by numerical examinations of the posterior estimation results.

The application for ISBE survey data showed that the estimates of model parameters and the small area statistics can differ significantly when different models are used. The illustration based on a simulated population showed that the use of improper models can produce seriously misleading results. Therefore, a careful model examination is crucial for producing reasonable and reliable estimation results. In addition, utilizing effective methods to choose appropriate models from the potentials at this early stage of data exploration is highly desirable.

In this chapter, we will focus on studying more efficient and effective model selection methods in Bayesian framework. Some existing methods for Bayesian model checking and model comparison will be introduced in Section 5.1. In Section 5.2, a method that benchmarks HB estimates with respect to higher level direct estimates and measures the relative inflation in the posterior mean squared error due to benchmarking in the posterior predictions is developed to evaluate the performance of hierarchical models. The performance of the proposed benchmarked hierarchical Bayesian posterior predictive model comparison method is examined using an illustrative example in Section 5.3. The method is then applied for model selection in the analysis of the actual ISBE survey data in Section 5.4. Section 5.5 will have some concluding remarks.

5.1 Existing methods for Bayesian model checking and model comparison

Model checking and model selection have always been important dimensions 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. Model checking is used to determine if a model provides an adequate fit to a given data set. There has been a broad exploration of model checking in the literature. Some methods compare estimates of parameters under a larger model to their corresponding null values under a smaller model. Some methods compare the observed data to what would be obtained under the assumed model based on visual examination of diagnostic plots (e.g., Gelman (2004)). Other methods compare the posterior distribution of a diagnostic function of data and/or parameters to its assumed prior distribution numerically. Model selection is very closely related to model checking except that the goal of model selection is to select the best of a set of models instead of assessing the goodness-of-fit of a single model. The appropriateness of a model is determined 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 Bayesian methods of model comparison and model selection rely on *Bayes factors* (Kass and Raftery (1995)). Bayes factors, proposed by Jeffreys in 1935, was developed for quantifying the evidence provided by the data in favor of a null hypothesis compared to an alternative hypothesis. It is defined as the ratio of the marginal likelihoods of data under the hypotheses, which is also the ratio of the posterior odds of the null hypothesis to its prior odds. When the prior probabilities of the hypotheses are both one-half, the Bayes factor is equal to the posterior odds of the null hypothesis. Bayes factors can be more widely used than the classic likelihood ratio tests because it can be used not only for the comparison between nested models but also for non-nested models. When a group of models are being considered, the Bayes factors can be used for obtaining the posterior probabilities of the models given the data. However, to use Bayes factors, it is necessary to specify proper prior distributions

for the parameters and models. It could be a lot of 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 distributions of parameters, which in general is not desirable especially in the preliminary stage of data analysis.

Alternatively, recent developments have been focussed on methods from a posterior predictive perspective. In this point of view, the observed data will be compared with what would be obtained from the posterior predictive distribution under the assumed model based on a certain criterion. An extreme pattern of the predictive data relative to the observed data indicates the incompatibility of the data and the assumed model. Best model is chosen as a succinct model from which the predictive data best mimic the observed data in terms of the specified criterion. Like the Bayes factors, the posterior predictive approaches can be used for comparison across a large class of non-nested models. In contrast, the method allows the utilization of objective (improper) prior distributions as long as the resulting posterior distributions of parameters are proper. Hence, one can avoid the trouble of needing to carefully specify proper prior distributions by incorporating external information, especially when such information is not guaranteed to be available and when we are in a preliminary stage of model exploration and much effort on prior specification could be a waste if the hypothesized model is not appropriate. Among the posterior predictive approaches, the posterior predictive p-value, the L-criterion, and the deviance information criterion are commonly used, which are reviewed in this section.

5.1.1 Posterior Predictive P-Value

As a well-known inferential tool in goodness-of-fit model checking, the p-value provides a measure of "surprise" of the data against a hypothesized (null) model. It is a ubiquitous measure of quantifying the incompatibility of the observed data and the null model based on evaluating the probability of observing data that is more extreme than the collected data under the assumed model. With a chosen statistic $T(\boldsymbol{y})$, where $\boldsymbol{y}$ is the vector of the data values, the

p-value calculates the tail-area probability corresponding to the observed value of the statistic. Without loosing generality, we assume the larger the value of $T(\boldsymbol{y})$ is, the greater the degree of departure the data shows from the assumed model. Otherwise, we can choose to use $-T(\boldsymbol{y})$.

In a Bayesian framework, several p-values have been proposed based on choosing different predictive distributions of the statistic. A general notation for theses p-values can be given by

$$p = Pr^{h(\cdot)}(T(\boldsymbol{Y}) \geq T(\boldsymbol{y}^{obs})) \tag{5.1}$$

where $h(\cdot)$ is some specified predictive distribution for the statistic $T(\boldsymbol{y})$. For example, the empirical Bayes (plug-in) p-value is given by using

$$h(\cdot) = f(t|\hat{\boldsymbol{\theta}}), \tag{5.2}$$

where $\boldsymbol{\theta}$ is a vector of model parameters and $\hat{\boldsymbol{\theta}}$ is the maximum likelihood (ML) or restricted maximum likelihood (REML) estimate of $\boldsymbol{\theta}$. The posterior predictive p-value uses

$$h(\cdot) = \int f(t|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{y}^{obs})d\boldsymbol{\theta}, \tag{5.3}$$

which gets rid of the model parameters by integrating out $\boldsymbol{\theta}$ with respect to its posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{y}^{obs})$. The partial posterior predictive p-value uses the partial posterior predictive distribution

$$h(\cdot) = \int f(t|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{y}^{obs}\backslash t^{obs})d\boldsymbol{\theta}, \tag{5.4}$$

where

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}^{obs}\backslash t^{obs}) \propto f(\boldsymbol{y}^{obs}|t^{obs},\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \propto f(\boldsymbol{y}^{obs}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})/f(t^{obs}|\boldsymbol{\theta}), \tag{5.5}$$

which updates the prior $\pi(\boldsymbol{\theta})$ with the information in $\boldsymbol{y}^{obs}$ not in t^{obs}. The $h(\cdot)$ for different predictive distributions might not have a closed form of expression, but it can be approximated using MCMC methods, and thus the corresponding p-value could be easily calculated using the simulated values from MCMC.

Bayarri and Castellanos (2007) argued that the partial posterior predictive p-value can avoid the double use of the data and has more power than the empirical Bayes and posterior predictive p-values. The empirical Bayes predictive distribution centers around the same

place as the posterior predictive distribution, but is more concentrated than the latter, thereby producing less conservative p-values. However, with the help of the well-developed statistical software for implementing MCMC schemes such as WinBUGS, the calculation of posterior predictive p-values has become much easier and very straightforward for either large or small sample cases. Also, with one set of MCMC simulation values, the p-values for multiple test statistics can be easily computed. In contrast, the derivation of the ML or REML estimates of parameters for small sample problems are usually problematic in the calculation of empirical Bayesian p-values. In addition, the measure of uncertainty associated with the empirical Bayes estimators, given by mean squared error (MSE), is usually more involved and less straightforward than the measure of uncertainty for hierarchical Bayesian estimators. So is the interval estimation. For calculating the partial posterior predictive p-value, some MCMC schemes are needed to be specially designed for a chosen test statistic for approximating the partial conditional distributions, which requires more work in programming to implement MCMC simulation than using well-developed software. Moreover, each MCMC procedure is designed for each specified test statistic. In a circumstance that an optimal test statistic is not easy to specify, the use of multiple test statistics might be needed and thus require multiple designs and programs for MCMC simulation, which will dramatically increase the work load of computation. In the ISBE survey, considering that small samples are presented within strata, which can cause a problem in obtaining the ML or REML estimates of parameters and a single optimal test statistic is hard to specify under such a complex model structure, we prefer not to use empirical Bayes and partial posterior predictive p-values and choose the more conservative posterior predictive p-value.

The posterior predictive p-value was introduced by Guttman (1967) and Rubin (1984), who proposed to calculate the tail-area probability corresponding to the observed value of a statistic by using the posterior predictive distribution of the statistic. Meng (1994) and Gelman et al. (1996) formalized and extended the posterior predictive assessment of model fitness by introducing parameter-dependent discrepancy measures.

The posterior predictive p-value is the probability that the data from the posterior predic-

tive distribution is more extreme than the observed data in terms of a discrepancy measure. The basic discrepancy measures are test statistics, such as means, percentiles, ranges and some ancillary statistics under the assumed model. For example, to check the dispersion of the distribution, one could use the range; to check the behavior of the right (or left) tail, one could use the maximum (or minimum). Let $\boldsymbol{y}$ and $\boldsymbol{\theta}$ denote the vectors of data and the unknown parameters respectively and $T(\boldsymbol{y})$ be a test statistic. The posterior predictive p-value based on $T(\boldsymbol{y})$ is

$$p_{post} = Pr\{T(\boldsymbol{y}^{pred}) > T(\boldsymbol{y}^{obs})|\boldsymbol{y}^{obs}\}. \tag{5.6}$$

The $\boldsymbol{y}^{obs}$ and $\boldsymbol{y}^{pred}$ stand for the actual observed data and the replicate predicted data values under the posterior predictive distribution:

$$f(\boldsymbol{y}^{pred}|\boldsymbol{y}^{obs}) = \int f(\boldsymbol{y}^{pred}|\boldsymbol{\theta}, \boldsymbol{y}^{obs})\pi(\boldsymbol{\theta}|\boldsymbol{y}^{obs})d\boldsymbol{\theta}, \tag{5.7}$$

where $\pi(\boldsymbol{\theta}|\boldsymbol{y}^{obs}) = f(\boldsymbol{y}^{obs}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})/\int f(\boldsymbol{y}^{obs}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$ is the posterior distribution of parameters given the observed data and $\pi(\boldsymbol{\theta})$ is the prior distribution of parameters.

More generally, the discrepancy measure could involve the unknown nuisance parameters from the model. Let $D(\boldsymbol{y}, \boldsymbol{\theta})$ represent a generalized discrepancy measure, a function of both data and parameters. The posterior predictive p-value based on $D(\boldsymbol{y}, \boldsymbol{\theta})$ is

$$p_{post} = Pr\{D(\boldsymbol{y}^{pred}, \boldsymbol{\theta}) > D(\boldsymbol{y}^{obs}, \boldsymbol{\theta})|\boldsymbol{y}^{obs}\}. \tag{5.8}$$

The probability is taken over the joint posterior distribution

$$f(\boldsymbol{y}^{pred}, \boldsymbol{\theta}|\boldsymbol{y}^{obs}) = f(\boldsymbol{y}^{pred}|\boldsymbol{\theta}, \boldsymbol{y}^{obs})\pi(\boldsymbol{\theta}|\boldsymbol{y}^{obs}). \tag{5.9}$$

Meng (1994) pointed out that the use of a generalized discrepancy variable is an important improvement because it allows us to measure directly the discrepancy between sample quantities and population quantities when checking the discrepancy between the data and the assumptions. One of the commonly used discrepancy measures that involves the nuisance parameters is the χ^2 discrepancy defined as

$$\mathrm{X}^2(\boldsymbol{y}; \boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{(y_i - E(y_i|\boldsymbol{\theta}))^2}{Var(y_i|\boldsymbol{\theta})}, \tag{5.10}$$

where i indexes the cases in the sample in a generic manner.

The calculation of the posterior predictive p-value is straightforward in MCMC simulation. In MCMC computation, replicated predictive values of parameters or data values are drawn from the approximated posterior predictive distribution. Then the posterior predictive p-value is 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 (Gelman et al. (1996)). Concretely, the posterior predictive p-value based on a generalized discrepancy measure $D(\boldsymbol{y}, \boldsymbol{\theta})$ can be calculated in the following three steps.

1. Generate R values $\boldsymbol{\theta}^r, r = 1, \cdots, R$, of $\boldsymbol{\theta}$ given the observed data from $\pi(\boldsymbol{\theta}|\boldsymbol{y}^{obs})$.

2. Generate replicate predicted data $\boldsymbol{y}^{pred,r}$ independently from $f(\boldsymbol{y}|\boldsymbol{\theta}^r)$, $r = 1, \cdots, R$.

3. Compute the proportion of times out of R that $D(\boldsymbol{y}^{pred,r}, \boldsymbol{\theta}^r)$ is greater than $D(\boldsymbol{y}^{obs}, \boldsymbol{\theta}^r)$, $r = 1, \cdots, R$.

The posterior predictive p-value allows the use of objective non-informative prior distributions, which could be desirable for model checking and comparison especially at the preliminary stage of model exploration. The method is straightforward and relatively easy to implement in many hierarchical models for which MCMC simulation can be performed. Based on one set of simulations, one can use multiple discrepancy measures to evaluate a model; each discrepancy measure yields an alternative Bayesian p-value.

On the other hand, a shortcoming of the posterior predictive p-value 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 p-value (Stern and Cressie (2000); Larsen and Lu (2007)). The partial posterior predictive p-value has been discussed earlier. The cross-validated posterior predictive samples are generated from the posterior predictive distribution by leaving out individual groups of data and thus avoid double use of data. The method also allows the use of objective prior distributions and multiple

discrepancy measures. There are some issues related to computation and multiplicity about this method need further exploration, which will be discussed in Chapter 6. However, if these issues could be appropriately addressed, the method could be an alternative for model checking in the ISBE survey. Cross-validated posterior predictive checking is not considered here but could be studied in the future.

5.1.2 L-Criterion

The L-criterion, proposed by Laud and Ibrahim in 1995, is another approach of model selection from the Bayesian predictive point of view. It is a criterion-based approach, which measures the performance of a model by evaluating expected posterior predictive errors. In their arguments, they considered the posterior predictive distribution of the predictive data conditioned on the observed data under an assumed model to be the distribution of the response values in an imaginary replicate experiment. The replicate experiment is assumed to done under the same conditions as the current study. They referred to the posterior predictive distribution as the *predictive density of a replicated experiment* (PDRE) in their paper. The L-criterion is then defined as the square root of the expected squared Euclidean distance of predictive values from the observed values in the posterior predictive distribution. A calibration of the criterion is also proposed for interpreting the relative magnitudes of the criterion values for various models.

Suppose models under consideration are indexed by $m \in \mathcal{M}$ and have parameters $\boldsymbol{\theta}^{(m)} \in \Theta^{(m)}$, where $\mathcal{M}$ is the model space and $\Theta^{(m)}$ is the parameter space for model m. Let $\boldsymbol{y} = (y_i, i = 1, \ldots, n)'$ denote the responses measured in the study. Let $\boldsymbol{y}^{pred} = (y_i^{pred}, i = 1, \ldots, n)'$ denote random variables measuring responses in the hypothetical replicated study. For a particular model m, define

$$\begin{aligned} L_m^2 &= E\{\left(\boldsymbol{y}^{pred} - \boldsymbol{y}^{obs}\right)' \left(\boldsymbol{y}^{pred} - \boldsymbol{y}^{obs}\right) | \boldsymbol{y}^{obs}, m\} \\ &= \sum_{i=1}^{n} [\{E\left(y_i^{pred} | \boldsymbol{y}^{obs}\right) - y_i\}^2 + Var\left(y_i^{pred} | \boldsymbol{y}^{obs}\right)]. \end{aligned} \tag{5.11}$$

The expectation is taken with respect to the posterior predictive density of a replicated exper-

iment under a given model m:

$$f(\boldsymbol{y}^{pred}|\boldsymbol{y}^{obs}, m) = \int f(\boldsymbol{y}^{pred}|m, \boldsymbol{\theta}^{(m)})\pi(\boldsymbol{\theta}^{(m)}|m, \boldsymbol{y}^{obs})d\boldsymbol{\theta}^{(m)}. \tag{5.12}$$

The density $\pi(\boldsymbol{\theta}^{(m)}|m, \boldsymbol{y}^{obs}) \propto \pi(\boldsymbol{\theta}^{(m)}|m)f(\boldsymbol{y}^{obs}|m, \boldsymbol{\theta}^{(m)})$ is the posterior distribution of $\boldsymbol{\theta}^{(m)}$ under model m given observed data. L_m^2 measures the closeness of the predictive data to the observed data accounting for the variability of the predictions (Laud and Ibrahim (1995)). Small values of L_m^2 indicate good models. Laud and Ibrahim's L-criterion is defined as

$$L_m = \sqrt{L_m^2}, \tag{5.13}$$

which is measured in the same scale as the response variable.

To quantify the uncertainty that is inherent in the criterion values, they calculated the standard deviation of the criterion with respect to the marginal distribution of the outcome variable $\boldsymbol{Y}$ under the model with the smallest criterion value. Their *calibration number* for the L-criterion is

$$S_{L_{m^*}} = [Var\{L_{m^*}(\boldsymbol{Y})\}]^{1/2}, \tag{5.14}$$

where m^* denotes the model with the smallest criterion value. The calibration number can be calculated by drawing Monte Carlo samples of the response variable $\boldsymbol{Y}$ from its marginal distribution and obtaining the L-criterion value for each sample of $\boldsymbol{Y}$. For better presentation of the relative magnitudes of the criterion values for different models, Hoeting and Ibrahim (1998) defined a *calibration comparison score* (CCS) as

$$\phi_m = \frac{L_m - L_{m^*}}{S_{L_{m^*}}}, \tag{5.15}$$

which measures the number of calibration units that a given model m is from the model m^* with the smallest criterion value. A simple model with a relatively small CCS, say less than 2, is preferred.

Laud and Ibrahim (1995) showed several advantages of the L-criterion over other well accepted existing model selection criteria such as AIC and BIC, which do not allow prior input of external information, rely heavily on asymptotic considerations and do not have calibrations. The L-criterion approach, like the posterior predictive checking, can compare

a variety of models including models that do not have nested structures, allows the use of objective non-informative prior distributions It, too, has the shortcoming of double use of the observed data: the prediction in the replicated experiment is based on the observed data, which are then also used to calculate the criterion value.

The L-criterion emphasizes the observations instead of the parameters. It only measures the "distance" between the observed and the predicted data values but not the discrepancy between the sample data and population quantities. The evaluation of the L-criterion value is based on comparison among models. A model with the smallest criterion value must be identified and the interpretation of model performance is based on comparing the relative magnitude of criterion to the smallest value. So the L-criterion, unlike the posterior predictive p-values, is only valuable for comparison between models but not for assessing the performance of a single model.

5.1.3 Deviance Information Criterion

The deviance information criterion (DIC) is a hierarchical modeling generalization of the frequentist Akaike's Information criterion (AIC). It is generally considered as a Bayesian analogue of AIC with wider applicability. The DIC is intended to describe the predictive ability of models based on measuring the posterior expected loss of predicting replicate data from the same mechanism that generated the observed data when assuming a loss function $\mathcal{L}(\boldsymbol{Y}, \tilde{\boldsymbol{\theta}}) = -2\log\{p(\boldsymbol{Y}|\tilde{\boldsymbol{\theta}}\}$. The idea is analogous to the frequentist model comparison criterion except the latter is based on sampling expectations and asymptotic arguments.

A skeleton concept in DIC is the "deviance", which is defined as

$$D(\boldsymbol{y}, \boldsymbol{\theta}) = -2\log\{p(\boldsymbol{y}|\boldsymbol{\theta})\} + 2\log\{f(\boldsymbol{y})\}, \tag{5.16}$$

where $\boldsymbol{y}$ is the vector of data values and $f(\boldsymbol{y})$ is some fully specified standardizing term that is a function of the data alone. For example, for the exponential family with $E(\boldsymbol{y}) = \mu(\boldsymbol{\theta})$, $f(\boldsymbol{y}) = p\{\boldsymbol{y}|\mu(\boldsymbol{\theta}) = \boldsymbol{y}\}$. Spiegelhalter et al. (2002) considered a quantity

$$d_{\Theta}\{\boldsymbol{y}, \boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}(\boldsymbol{y})\} = -2\log\{p(\boldsymbol{y}|\boldsymbol{\theta})\} + 2\log[p\{\boldsymbol{y}|\tilde{\boldsymbol{\theta}}(\boldsymbol{y})\}], \tag{5.17}$$

where $\tilde{\boldsymbol{\theta}}(\boldsymbol{y})$ is an estimate of $\boldsymbol{\theta}$ given data and is generally taken to be the posterior mean of parameters $E(\boldsymbol{\theta}|\boldsymbol{y}^{obs}) = \bar{\boldsymbol{\theta}}$. This quantity is a measure of the true over the estimated residual information, and can be thought of as the reduction in surprise due to estimation. Then a Bayesian measure of model complexity is defined as the posterior expectation of this quantity:

$$\begin{aligned} p_D &= E_{\boldsymbol{\theta}|\boldsymbol{y}}[d_\Theta\{\boldsymbol{y}, \boldsymbol{\theta}, \bar{\boldsymbol{\theta}}(\boldsymbol{y})\}] \\ &= E_{\boldsymbol{\theta}|\boldsymbol{y}}[-2\log\{p\left(\boldsymbol{y}|\boldsymbol{\theta}\right)\}] + 2\log[p\{\boldsymbol{y}|\bar{\boldsymbol{\theta}}(\boldsymbol{y})\}] \\ &= \overline{D\left(\boldsymbol{y}, \boldsymbol{\theta}\right)} - D\left(\boldsymbol{y}, \bar{\boldsymbol{\theta}}\right), \end{aligned} \tag{5.18}$$

where $\overline{D\left(\boldsymbol{y}, \boldsymbol{\theta}\right)}$ is the posterior mean deviance and $D\left(\boldsymbol{y}, \bar{\boldsymbol{\theta}}\right)$ is the deviance at the posterior mean of parameters. The posterior mean deviance, $\overline{D\left(\boldsymbol{y}, \boldsymbol{\theta}\right)}$, as a measure of predictive accuracy is a Bayesian measure of overall model adequacy, which from equation (5.18) is equal to the "plug-in" measure of fit $D\left(\boldsymbol{y}, \bar{\boldsymbol{\theta}}\right)$ plus a measure of complexity p_D.

The computation of the expected deviance and the effective number of parameters are trivial by using MCMC methods. Suppose R draws of parameters from L chains are retained for posterior estimation. With $L \cdot R$ draws of $\boldsymbol{\theta}$ in total from its posterior distribution, the posterior mean deviance is then calculated by

$$\hat{\overline{D\left(\boldsymbol{y}, \boldsymbol{\theta}\right)}} = \frac{1}{LR}\sum_{l=1}^{L}\sum_{r=1}^{R} D\left(\boldsymbol{y}^{obs}, \theta^{(lr)}\right), \tag{5.19}$$

where l denotes the chain and r the iteration. The model complexity, the effective number of parameters, of a Bayesian model is then obtained by

$$\hat{p}_D = \hat{\overline{D\left(\boldsymbol{y}, \boldsymbol{\theta}\right)}} - D\left(\boldsymbol{y}, \hat{\bar{\boldsymbol{\theta}}}\right), \tag{5.20}$$

where $D\left(\boldsymbol{y}, \hat{\bar{\boldsymbol{\theta}}}\right)$ is the deviance at the estimate of the posterior mean of $\boldsymbol{\theta}$.

The classical AIC criterion for model comparison is generally based on a "plug-in" estimate of fit plus twice the effective number of parameters in the model. By analogy with the classic result, the Bayesian criteria for model comparison is then proposed as a Bayesian "plug-in" estimate of fit plus twice the effective number of parameters:

$$\begin{aligned} \text{DIC} &= D\left(\boldsymbol{y}, \bar{\boldsymbol{\theta}}\right) + 2p_D \\ &= \overline{D\left(\boldsymbol{y}, \boldsymbol{\theta}\right)} + p_D, \end{aligned} \tag{5.21}$$

which is referred to as the deviance information criterion (DIC; Spiegelhalter et al. 2002). It can be considered as a Bayesian measure of adequacy $\overline{D(\boldsymbol{y}, \boldsymbol{\theta})}$ penalized by the model complexity p_D.

In general, the posterior mean deviance tends to decrease as the number of parameters in the model increases. On the other hand, the DIC takes the dimension of effective parameters into account, which usually compensates for this effect in favor of models with a smaller number of parameters, and therefore provides a combined better overall evaluation of model performance. The model with the smallest DIC is estimated to be the model that would best predict a replicate data set which has the same structure as that is currently observed. A large difference of DIC values, say bigger than 10, would provide a sufficient evidence in favor of the model with the smaller DIC value. However, if only a fairly small difference of DIC values is observed, then it might not be appropriate to choose the model with the lowest DIC value, especially when the models make very different inferences. The decision should be based on more careful examination of other characteristics of the models.

Spiegelhalter et al. (2002) showed that in models with approximately normal likelihoods and negligible prior information, the effective number of parameters is approximately the true number of parameters and thus the DIC will be approximately equivalent to the AIC. The DIC provides a way of incorporating external information and also allows the use of objective non-informative priors. The implementation of DIC as part of MCMC estimation in most hierarchical models is very convenient and can be done using WinBUGS. However, for some models such as mixture models which are commonly used for missing data problems and problems with complex population structures, DIC is not so straightforward to implement. Since the observed data are used both to construct the posterior distribution and to compute the posterior mean deviance, the method is also conservative in detecting model failure to some degree due to the double use of the observed data.

An alternative effective number of parameters was proposed in Gelman et al. (2004). Consider a non-hierarchical model with weak prior information, the posterior variance of the deviance is approximately twice the number of parameters in the model. So in models with

negligible prior information, half the variance of the deviance is a reasonable estimate of the effective number of parameters in a model. This might suggest using

$$p_V = \mathrm{Var}_{\boldsymbol{\theta}|\boldsymbol{y}}\{D(\boldsymbol{y}, \boldsymbol{\theta})\}/2 \tag{5.22}$$

as an estimate of the effective number of parameters in a model in more general situations. The p_V has shown to be invariant to parameterisation, generally more robust than p_D and easy to the calculate. When there is moderate shrinkage in the model, p_V is usually expected to be larger than p_D. The model complexity p_V can be estimated from the posterior simulations by

$$\hat{p}_V = \frac{1}{2(LR-1)} \sum_{l=1}^{L} \sum_{r=1}^{R} \left(D(\boldsymbol{y}^{obs}, \theta^{(lr)}) - \overline{\hat{D(\boldsymbol{y}, \boldsymbol{\theta})}} \right)^2. \tag{5.23}$$

5.2 Benchmarked HB Model Comparison

Sample surveys are usually used for obtaining not only estimates for whole populations but also estimates for subpopulations. However, due to restrictions of various sorts, surveys are often designed to ensure reliable direct estimates in only large regions. In many studies in which small areas are of interest as well, it can happen that only small sample sizes are allocated to individual small areas. As a result, direct survey estimates are no longer able to produce reliable estimates due to small sample sizes. Some indirect estimates based on implicit or explicit model assumptions are needed in order to "borrow strength" from related areas to increase the effective sample sizes and consequently the precision and reliability of the estimation.

When model-based estimators are used, it is often desirable to have a thorough model examination or at least some sort of "calibration" of the result. Particularly, in studies in which small areas are of interest but have small sample sizes, when there is a big enough sample for producing reliable estimates for larger regions composed of groups of small areas, benchmarking the model-based estimators to reliable direct estimates for large areas is desirable to protect against possible model mis-specification. You et al. (2004) suggested to "benchmark" the Hierarchical Bayesian (HB) estimators of small areas so that the benchmarked HB estimators

will add up to the direct estimators in larger regions and proposed a measure of uncertainty for the benchmarked HB estimator.

5.2.1 Benchmarked HB estimator

In the Iowa EP survey, the whole state was stratified by two factors that are crossed to define strata: district size and AEA. The populations of small, medium and large size levels and the populations of individual AEAs are higher levels of aggregation of small areas (strata). The idea of benchmarking can be applied for the whole state, for the areas aggregated by size, for the areas aggregated by AEA, or in two dimensions of aggregation by size and by AEA. For example, we can benchmark the HB estimators for AEAs (strata) in a certain size level so that the sum of the *benchmarked HB (BHB)* 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

$$\sum_j N_{ij}\hat{\mu}_{ij}^{BHB} = \sum_j N_{ij}\hat{\bar{y}}_{ij}, \tag{5.24}$$

where $i \in \{\text{size level: } 1 = \text{large}; 2 = \text{medium}; 3 = \text{small}\}$, $j \in \{12 \text{ AEAs}\}$, $\hat{\mu}_{ij}^{BHB}$ and $\hat{\bar{y}}_{ij}$ denote the benchmarked HB and the direct estimators of the population mean for stratum (i, j). In particular, the *raking-benchmarked HB* (RBHB) estimator for stratum (i, j) is given by

$$\hat{\mu}_{ij}^{RBHB} = \hat{\mu}_{ij}^{HB} \frac{\sum_j N_{ij}\hat{\bar{y}}_{ij}}{\sum_j N_{ij}\hat{\mu}_{ij}^{HB}}. \tag{5.25}$$

The $\hat{\mu}_{ij}^{HB}$ is the HB estimator for stratum (i, j).

The variation associated with the BHB estimator under the assumed model can be measured by the *posterior mean squared error* (PMSE):

$$\text{PMSE}(\hat{\mu}_{ij}^{BHB}) = E[(\hat{\mu}_{ij}^{BHB} - \mu_{ij})^2 | \boldsymbol{y}^{obs}]. \tag{5.26}$$

Since $E[(\hat{\mu}_{ij}^{HB} - \mu_{ij})|\boldsymbol{y}^{obs}] = 0$ under the assumed model, the PMSE of the HB estimator is actually equal to the posterior variance $V\left(\mu_{ij}|\boldsymbol{y}^{obs}\right)$:

$$\begin{aligned}
\text{PMSE}(\hat{\mu}_{ij}^{BHB}) &= E[(\hat{\mu}_{ij}^{HB} - \mu_{ij})^2 | \boldsymbol{y}^{obs}] \\
&= V\left(\mu_{ij}|\boldsymbol{y}^{obs}\right) + \{E[(\hat{\mu}_{ij}^{HB} - \mu_{ij})|\boldsymbol{y}^{obs}]\}^2 \\
&= V\left(\mu_{ij}|\boldsymbol{y}^{obs}\right).
\end{aligned}$$

By expressing the squared difference in the posterior expectation as the sum of three components,

$$\begin{aligned} & E[(\hat{\mu}_{ij}^{BHB} - \mu_{ij})^2 | \boldsymbol{y}^{obs}] \\ = \; & E[(\hat{\mu}_{ij}^{BHB} - \hat{\mu}_{ij}^{HB} + \hat{\mu}_{ij}^{HB} - \mu_{ij})^2 | \boldsymbol{y}^{obs}] \\ = \; & E[(\hat{\mu}_{ij}^{BHB} - \hat{\mu}_{ij}^{HB})^2 | \boldsymbol{y}^{obs}] + 2E[(\hat{\mu}_{ij}^{BHB} - \hat{\mu}_{ij}^{HB})(\hat{\mu}_{ij}^{HB} - \mu_{ij}) | \boldsymbol{y}^{obs}] + E[(\hat{\mu}_{ij}^{HB} - \mu_{ij})^2 | \boldsymbol{y}^{obs}], \end{aligned}$$

and noting that the cross-product term is indeed 0:

$$\begin{aligned} & E[(\hat{\mu}_{ij}^{BHB} - \hat{\mu}_{ij}^{HB})(\hat{\mu}_{ij}^{HB} - \mu_{ij}) | \boldsymbol{y}^{obs}] \\ = \; & (\hat{\mu}_{ij}^{BHB} - \hat{\mu}_{ij}^{HB}) E[(\hat{\mu}_{ij}^{HB} - \mu_{ij}) | \boldsymbol{y}^{obs}] = 0, \end{aligned}$$

the PMSE of the BHB estimator can be estimated as

$$\text{PMSE}\left(\hat{\mu}_{ij}^{BHB}\right) = V\left(\mu_{ij} | \boldsymbol{y}^{obs}\right) + \left(\hat{\mu}_{ij}^{BHB} - \hat{\mu}_{ij}^{HB}\right)^2, \tag{5.27}$$

which is the posterior variance $V\left(\mu_{ij} | \boldsymbol{y}^{obs}\right)$, plus a bias correction term $\left(\hat{\mu}_{ij}^{BHB} - \hat{\mu}_{ij}^{HB}\right)^2$.

The benchmarked HB estimators are design consistent in larger regions, which is an attractive property to survey practitioners. 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 bias the model-based estimator has, and thus the more serious inflation of PMSE there could be. Below we describe using benchmarked HB estimation results and PMSE inflation for the purpose of model selection.

5.2.2 The use of benchmarked discrepancy in posterior predictive model selection

From formula (5.27) for calculating the PMSE of the BHB estimator, we notice that the worse the data is fitted by the model, the larger the inflation of the PMSE due to benchmarking the HB estimator generally tends to be. Therefore a big inflation of PMSE can be a possible suggestion of model inadequacy. We consider measuring the discrepancy between the model

and the observed data based on measuring the degree to which benchmarking to reliable direct estimates for large areas inflates the PMSE of the HB estimator for small areas.

Let

$$\Delta = [\text{PMSE}(\hat{\mu}^{BHB}) - \text{PMSE}(\hat{\mu}^{HB})]/\text{PMSE}(\hat{\mu}^{HB}), \tag{5.28}$$

which is the relative change of the PMSE of the BHB to the HB estimator due to benchmarking. Since $\text{PMSE}(\hat{\mu}^{HB}) = E[(\hat{\mu}^{HB} - \mu)^2 | \boldsymbol{y}^{obs}]$ and $\text{PMSE}(\hat{\mu}^{BHB}) = (\hat{\mu}^{BHB} - \hat{\mu}^{HB})^2 + \text{PMSE}(\hat{\mu}^{HB})$, equivalently

$$\Delta = (\hat{\mu}^{BHB} - \hat{\mu}^{HB})^2 / E[(\hat{\mu}^{HB} - \mu)^2 | \boldsymbol{y}^{obs}]. \tag{5.29}$$

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 a discrepancy $D_1^{BHB}(\boldsymbol{y}, \boldsymbol{\theta})$ as the proportion of small areas having Δ_h bigger than a certain threshold value δ:

$$D_1^{BHB}(\boldsymbol{y}, \boldsymbol{\theta}) = H^{-1} \sum_{h=1}^{H} I_{\Delta_h > \delta}, \tag{5.30}$$

where

$$I_{\Delta_h > \delta} = \begin{cases} 1 & \text{if } \Delta_h > \delta \\ 0 & \text{otherwise} \end{cases} \tag{5.31}$$

and H is the total number of small areas in the population. 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 $|\hat{\mu}_h^{BHB} - \hat{\mu}_h^{HB}| > z_{0.975}\sqrt{V(\mu_h | \boldsymbol{y}^{obs})}$, 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 magnitude of inflation of the PMSE for the BHB versus the HB as

$$D_2^{BHB}(\boldsymbol{y}, \boldsymbol{\theta}) = H^{-1} \sum_{h=1}^{H} \Delta_h, \tag{5.32}$$

which is the average relative change of PMSE over all small areas.

For a given discrepancy $D(\boldsymbol{y}, \boldsymbol{\theta})$, the posterior predictive check will be based on the comparison between the predictive discrepancy $D(\boldsymbol{y}^{pred}, \boldsymbol{\theta})$ and the realized discrepancy $D(\boldsymbol{y}^{obs}, \boldsymbol{\theta})$.

The posterior predictive p-values based on the discrete discrepancy $D_1^{BHB}(\boldsymbol{y}, \boldsymbol{\theta})$ and on the continuous discrepancy $D_2^{BHB}(\boldsymbol{y}, \boldsymbol{\theta})$ are defined as

$$p_{post,1}^{BHB} = Pr(D_1^{BHB}(\boldsymbol{y}^{pred}, \boldsymbol{\theta}) \geq D_1^{BHB}(\boldsymbol{y}^{obs}, \boldsymbol{\theta}) | \boldsymbol{y}^{obs}) \tag{5.33}$$

and

$$p_{post,2}^{BHB} = Pr(D_2^{BHB}(\boldsymbol{y}^{pred}, \boldsymbol{\theta}) > D_2^{BHB}(\boldsymbol{y}^{obs}, \boldsymbol{\theta}) | \boldsymbol{y}^{obs}). \tag{5.34}$$

The estimation of these posterior predictive discrepancies can be accomplished from MCMC output with a little effort. In particular, for each value of $\boldsymbol{\theta}$ and $\boldsymbol{y}^{pred}$, one must compute $D_1^{BHB}(\boldsymbol{y}, \boldsymbol{\theta})$ and $D_2^{BHB}(\boldsymbol{y}, \boldsymbol{\theta})$. The computation could be large, because both the HB and BHB estimates themselves are computed using MCMC for each $\boldsymbol{y}^{pred}$. However, by using loops, not much extra programming effort is needed. Considering the speed and the data storage capacity of the modern computers, the computation absolutely can be realized with no difficulty.

5.3 Illustration

To illustrate the performance of the proposed model comparison method, we conduct model selection using the simulated finite population in Section 4.4.

Recall the finite population of EP courses taken by twelfth grade students from Iowa's public high schools which was simulated from a Poisson log-linear model:

$$\begin{aligned} y_{ijkl}|\lambda_{ijk} &\sim \text{Poisson } (\lambda_{ijk}) \\ \log(\lambda_{ijk}) &= \beta_0 + \beta_1 x_{ijk;1} + \tau_i + \eta_j. \end{aligned}$$

The number of EP courses taken by the $l-$th student in the $k-$th school in stratum (i, j), y_{ijkl}, follows a Poisson distribution with mean λ_{ijk}, where λ_{ijk} denotes the underlying school-specific rate of taking EP courses. 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 log-rate of taking EP courses is then assumed to be linearly related to some covariable variable at the school and the random effects from size levels and AEAs.

The enrollment size in twelfth grade was used as an auxiliary variable $\boldsymbol{x}_1$ in generating the population data set. The τ_i, $i = 1, \cdots, I = 3$ and η_j, $j = 1, \cdots, J = 12$ are the size and AEA random effects. The values of the parameters used for generating the population were spell out in Section 4.4. 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 using the stratified three-stage design described in Chapter 2.

Several candidate models for $\log(\lambda_{ijk})$, reflecting different assumptions about the between-school variation in log-rate of taking EP courses and the involvement of covariable variables, were considered and fitted to the sample data. The models are described in Table 5.1. Models 1-4 are a series of nested models, which all assume a linear relationship with the school enrollment size $x_{ijk;1}$. Model 1 assumes no random effects. Model 2 allows size correlated variation between the school-specific rates. Models 3-4 include random effects from size levels and AEAs with and without interactions. Model 5 assumes random effects from both size and AEAs but no relationship with covariable variables. Model 6 considers only a relationship with covariate variables but no variation between the school-specific rates. Model 7 adds extra individual school variation to the log-rates of schools. The covariate variables $x_{ijk;q}, q = 1, \cdots, 5$ in model 6 correspond to auxiliary information at the school level 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 parameters $\beta_q, q = 1, \cdots, 5$ are the regression coefficients of the covariate variables. Among these models, model 3 is the model from which the population was simulated.

The prior distributions for model parameters are assumed to be mutually independent. A (proper but weak) normal prior with very large variance is specified for the intercept β_0. Improper uniform priors are placed on the coefficient parameters β_q, $q = 1, \cdots, 5$; The random effects τ_i, η_j and ζ_{ij} are all assumed to have normal prior distributions having zero mean and precisions σ_τ^{-2}, σ_η^{-2} and σ_ζ^{-2}. Weakly informative priors Gamma$(0.001, 0.001)$ are assumed for the random effects precision parameters. For each model, by running MCMC simulation in

Table 5.1 Seven models reflecting different assumptions about the between-school variation in log-rate of taking EP courses and the involvement of covariable variables in the illustrative example.

Models	Log linear model equation
M1	$\log(\lambda_{ijk}) = \beta_0 + \beta_1 x_{ijk;1}$
M2	$\log(\lambda_{ijk}) = \beta_0 + \beta_1 x_{ijk;1} + \tau_i$
M3	$\log(\lambda_{ijk}) = \beta_0 + \beta_1 x_{ijk;1} + \tau_i + \eta_j$
M4	$\log(\lambda_{ijk}) = \beta_0 + \beta_1 x_{ijk;1} + \tau_i + \eta_j + \zeta_{ij}$
M5	$\log(\lambda_{ijk}) = \beta_0 + \tau_i + \eta_j$
M6	$\log(\lambda_{ijk}) = \beta_0 + \sum_{q=1}^{5} \beta_q x_{ijk;q}$
M7	$\log(\lambda_{ijk}) = \beta_0 + \beta_1 x_{ijk;1} + \tau_i + \eta_j + v_{ijk}$

WinBUGS, we independently simulated $L = 3$ parallel Markov chains, each of length $20,000$ iterations. The first 10,000 iterations for each chain are deleted as a "burn-in" period. By thinning to every 10^{th} iteration, 1,000 iterates from each chain are retained for posterior estimation.

Table 5.2 shows the results of comparing the seven models using the three methods discussed in Section 5.1 and the method of benchmarked HB posterior predictive checking proposed in Section 5.2. 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 discrepancies (5.30) and (5.32) are denoted $p^{BHB}_{post,1}$ and $p^{BHB}_{post,2}$, respectively. A very small p-value, say less than .05, shows extreme patterns of the observed data relative to the posterior predictive data generated from the fitted model in terms of the specified discrepancy measure. Hence, small p-values indicate evident incompatibility between data and the fitted model. The "CCS" stands for the calibration comparison score (ϕ_m) for the L-criterion value (L_m). A relatively small CCS, say less than 2, places no strong evidence against the model. The "DIC" represents the deviance information criterion value. The smallest DIC value suggests a best model. The p_D denotes the effective number of parameters, which is Bayesian measure of model complexity.

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. Among the acceptable models, model 3 is the most parsimonious. When using the L-criterion, model 7

has the smallest L_m value. Calibrated by the standard deviation of the criterion value under model 7, the calibration comparison scores (CCSs; ϕ_ms) for models 1, 2, and 6 are larger than the value of 3 and are too big in terms of calibration of inherent variation of criterion values. Model 5 is the most succinct model with a not extreme CCS. Among models having small DIC values, model 3 is the simplest model. When using benchmarked HB model selection based on discrepancy $D_1^{BHB}(\boldsymbol{y}, \boldsymbol{\theta})$, 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 $D_2^{BHB}(\boldsymbol{y}, \boldsymbol{\theta})$, which measures the degree of inflation of PMSE quantitatively instead of simply counting the amount of small areas having extreme inflation of PMSE, 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 $D_2^{BHB}(\boldsymbol{y}, \boldsymbol{\theta})$. Model 5 is the simplest model among models with $p_{post,2}^{BHB}$ bigger than 0.05. As a summary, 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. Considering the fact that the coefficient parameter of $\boldsymbol{x}_1$ is given a very small value in the simulation in order to match the magnitude of the actual survey data and the range of $\boldsymbol{x}_1$ (the logarithm of enrollment size) is also very short so that the effect of the first covariate term is small relative to other effects, models 3 and 5 could be indeed indistinguishable in terms of making inferences and predictions about the stratum means. The $p_{post,1}^{BHB}$ fails to detect the significant model inadequacy of model 1. The reason could be that the discrepancy $D_1^{BHB}(\boldsymbol{y}, \boldsymbol{\theta})$, being discrete, probably loses some power relative to the more quantitative $D_2^{BHB}(\boldsymbol{y}, \boldsymbol{\theta})$.

In addition, we compare the HB estimators under models 3 and 5 in terms of the absolute relative bias (ARB), which is defined as the absolute value of the relative bias of the estimate over the realized finite population value, and the posterior variance (the same as the PMSE). Model 5 produces slightly larger ARB to the realized finite population mean and a little bit higher posterior variance in most of the strata. The HB estimates of stratum means under these two models, however, are very similar. Basically, all the Bayesian model comparison

Table 5.2 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 (ϕ_m) for the L_m statistic. DIC = deviance information criterion. p_D = effective number of parameters. $p_{post}^{1;BHB}$ = posterior predictive p-value based on the discrepancy $D_1^{BHB}(\boldsymbol{y}, \boldsymbol{\theta})$. $p_{post,2}^{BHB}$ = posterior predictive p-value based on the discrepancy $D_2^{BHB}(\boldsymbol{y}, \boldsymbol{\theta})$. Bold values indicate models that can 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

methods discussed above except the posterior predictive checking based on the discrepancy $D_1^{BHB}(\boldsymbol{y}, \boldsymbol{\theta})$ worked 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.

Below are some results from the illustrative example showing how the HB estimates are affected by the appropriateness of model assumptions and how benchmarking effects estimates under the correct and incorrect models.

The HB estimator can "borrow strength" from related areas through hierarchical modeling and produce more reliable estimates of small areas than direct estimator. However, just like all model-based estimators, the HB estimator is also vulnerable to mis-specification of models. Figure 5.1 shows the absolute relative biases (ARB) of HB estimates over the realized (true) finite population mean for 12 strata of medium districts, when the true (correct) model and when an inadequate model are used. The strata are sorted by the population size of PSUs. Larger strata get more PSUs in the sample. The five strata on the left hand side have one PSU sampled and the seven strata on the right hand side have two PSUs sampled. For the single

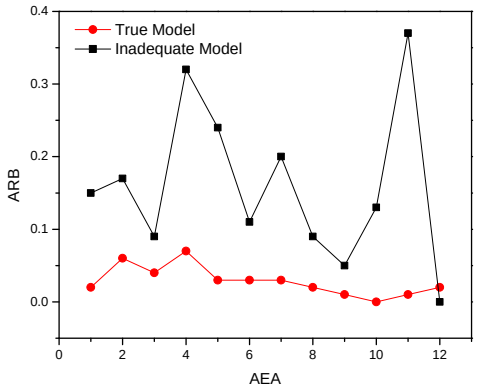


Figure 5.1 The absolute relative biases (ARBs) of HB estimates under the true model (model 3) and an inadequate model (model 2) in the illustrative example.

randomly selected sample, when the assumed model is correct, 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. However, when a smaller model, which fails to address the random effect from AEAs and had shown strong evidence of model inadequacy in the previous model checking, is used, the ARBs for HB estimates are much bigger than those under the true model. The largest ARB is near 40%. Four strata have ARBs more than 20% and the majority is above or around 10%. Figure 5.2 shows the ARBs of BHB estimates under the true and the smaller models, which are fairly close in our case. This indicates that the BHB estimates are robust to model failure to a certain degree.

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 the BHB estimator could be extremely large. Figures 5.3 display the RPMSEs of the HB and the BHB estimators under the true model and the smaller model. 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 inadequate model is much bigger than using the true model. This is because model failure caused a serious bias of the HB estimator, which still has a large estimate of precision, and correspondingly a big bias correction term for the PMSE of the BHB estimator.

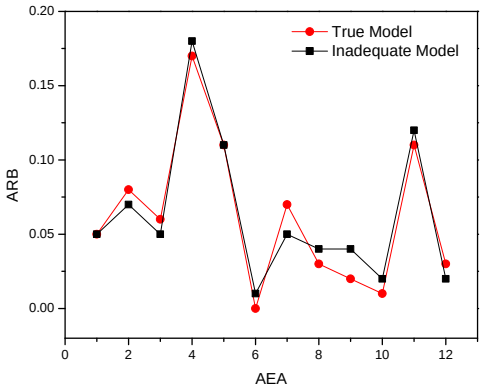


Figure 5.2 The absolute relative biases (ARBs) of BHB estimates under the true model (model 3) and an inadequate model (model 2) in the illustrative example.

Therefore, serious inflation of the PMSE of the BHB relative to the HB estimator could be an indication of a poorly specified model.

From the above figures, we see great advantages of using a HB 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 with an assessment of high precision. 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. Big inflation of the PMSE for the BHB over the HB estimator could be a "signal" of potential model failure. Quantifying the inflation of the PMSE could be used as a discrepancy measure in mode checking.

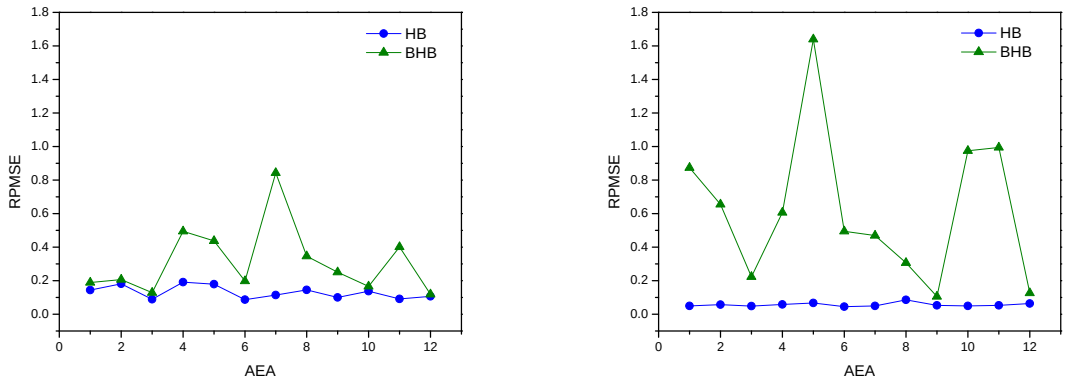


Figure 5.3 The root of posterior mean square error (RPMSE) of the HB and the BHB estimates under the true model (model 3) on the left and under an inadequate model (model 2) on the right in the illustrative example.

5.4 Model Selection in the Analysis of the ISBE Survey Data

The actual ISBE survey data were collected in 2005 from 51 sample schools in 11 AEAs. AEA 9 did not participate. AEA 1 had sample schools from only medium size districts; large and small districts were missing. So there were only 26 strata with sample schools in the actual data. All large districts and one medium district in AEA 1 had more than one schools. The rest of the districts (medium or small) have only one school each. In strata with medium districts, six strata had a single school in the sample. In strata with small districts, three strata had only one school sampled. More descriptions of the survey data could be found in Section 4.5.

Both design-based estimation and hierarchical Bayesian (HB) estimation were applied to analyze the survey data. The Horvitz-Thompson (HT) and ratio estimators were used in design-based estimation. The variances of estimators for strata with at least two PSUs sampled were estimated using direction variance estimation for a stratified multi-stage design. The variances of estimators for one-PSU strata were estimated using collapsed strata variance estimator followed by synthetic variance redistribution (CSSV) and collapsed strata restricted generalized variance function synthetic variance (CRGVFSV) estimation methods. The esti-

mators and variance estimation methods showed similar properties as they did in the previous simulation study (Lu and Larsen (2006, 2007)). More details about the results of design-based estimation were shown in Chapter 3. In hierarchical Bayesian analysis, we used a few generalized linear mixed models to analyze the survey data. The details about different models and estimators and the results of HB estimation were discussed in Chapter 4. In this section, we will focus on model comparison using the methods discussed in Section 5.1 and the method of posterior predictive checking using benchmarking proposed in Section 5.2 combined with some graphical examination methods.

There were also some other data sources from the Iowa Department of Education and National Center for Education Statistics (NCES) that are assessable online. The Common Core Data (CCD) from NCES contain general information on geography, administration and finance for public high schools, school districts and AEAs in Iowa. They also have student information such as the membership counts for different grades, demographic groups, and minority groups such as students with IEPs, students having free or reduced price lunch, immigrants, and English language learners. Staffing information is available as well. After examining the external administrative variables from these data sets, we chose over 30 variables as potential useful auxiliary variable for data analysis, which include the enrollment in grade nine to grade twelve and within each grade, the population size of different demographic groups and a variety of minority groups, the revenues and expenditures in different sources, and the counts of FTE teachers in various duty categories. However, we found that the variables are mostly highly correlated. So we conducted a simple factor analysis and found that the primary factor is the school enrollment which is dominating. Other factors related to the minority groups of Hispanic and Black students and students having free and reduced price lunch did not contribute very much to explaining the variability associated with the response variable, and thus would not be considered in further model (and variable) selection. The methods proposed should be able to select models using more predictor variables such as the Common Core variables in other education studies.

In the HB estimation, we fitted both Poisson-Lognormal and Poisson-Gamma models to the

survey data. We compared models with different involvement of auxiliary variables and random effects using Bayesian model comparison methods discussed above. Neither the calibration comparison scores (CCSs; ϕ_ms) using the L-criterion nor the DIC values indicated even a slight difference between the models applied to the data.

Table 5.3 Discrepancy measures used in posterior predictive checking for candidate models applied to the Iowa survey data. In the measures, $\omega_{ijk} = \sum_{l \in s_{ijk}} \omega_{ijkl}$.

D_1	χ^2 student	$\sum_i \sum_j \sum_k \sum_{l \in s_{ijk}} \frac{(y_{ijkl} - \omega_{ijkl}\lambda_{ijk})^2}{\omega_{ijkl}\lambda_{ijk}}$
D_2	χ^2 school	$\sum_i \sum_j \sum_k \frac{(\bar{y}_{ijk} - \gamma_{ijk})^2}{\gamma_{ijk}^2/\alpha}$ under Poisson-Gamma models $\sum_i \sum_j \sum_k \frac{(\bar{y}_{ijk} - \mu_{ijk})^2}{\mu_{ijk}^2\{\exp(\sigma_v^2) - 1\}}$ under Poisson-Lognormal models, where $\mu_{ijk} = \exp\left(\boldsymbol{\beta}'\boldsymbol{x}_{ijk} + \tau_i + \eta_j + \zeta_{ij} + \sigma_v^2/2\right)$ when random effects from size levels, AEAs and interactions are presented
D_3	χ^2 overall	$\sum_i \sum_j \sum_k \frac{(\bar{y}_{ijk} - \gamma_{ijk})^2}{\gamma_{ijk}/\omega_{ijk} + \gamma_{ijk}^2/\alpha}$ under Poisson-Gamma models $\sum_i \sum_j \sum_k \frac{(\bar{y}_{ijk} - \mu_{ijk})^2}{\mu_{ijk}/\omega_{ijk} + \mu_{ijk}^2\{\exp(\sigma_v^2) - 1\}}$ under Poisson-Lognormal models
D_4	maximum	$\max_{i,j,k,l \in s} (y_{ijkl}/\omega_{ijkl})$
D_5	negative minimum	$-\min_{i,j,k,l \in s} (y_{ijkl}/\omega_{ijkl})$
D_6	range	$\max_{i,j,k,l \in s} (y_{ijkl}/\omega_{ijkl}) - \min_{i,j,k,l \in s} (y_{ijkl}/\omega_{ijkl})$
D_7	maximum of school means	$\max_{i,j,k \in s} \bar{y}_{ijk}$
D_8	negative minimum of school means	$-\min_{i,j,k \in s} \bar{y}_{ijk}$
D_9	range of school means	$\max_{i,j,k \in s} \bar{y}_{ijk} - \min_{i,j,k \in s} \bar{y}_{ijk}$
D_{10}	maximum of school ranges	$\max_{i,j,k \in s}[\max_{l \in s} (y_{ijkl}/\omega_{ijkl}) - \min_{l \in s} (y_{ijkl}/\omega_{ijkl})]$

Table 5.3 lists ten discrepancy measures not including the new benchmarking ones that were used in posterior predictive checks of the models. D_1 is the χ^2 discrepancy of students within schools, which describes the fitness of the first level (Poisson) model. D_2 is the χ^2 discrepancy of schools, which describes the fitness of the second level (Gamma or Lognormal) model. D_3 is the overall χ^2 discrepancy for the hierarchical model. D_4 and D_5 are the maximum and (negative) minimum of rates of taking EP courses for students in the entire population, which

describe the left and right tail behaviors of the entire population distribution. D_6 is the range of rates for students, which also describes the dispersion of the hierarchical model. D_7 and D_8 are the maximum and (negative) minimum of school means, which describe the left and right tail behaviors of the second level model. D_9 is the range of school means, which describes the dispersion of the second level model. D_{10} is the maximum of school ranges, which also describes the overall performance of the hierarchical model. Measures D_5 and D_8 have negative signs because we want to choose the test statistics such that larger values of them indicate more discrepancy between the assumed model and the observed data.

Table 5.4 gives the posterior predictive p-values based on the discrepancy measures described in Table 5.3 for all candidate models. Here, we did not examine as many Poisson-Lognormal models as Poisson-Gamma models because we thought it could be adequate to reveal the relative properties of the two families of the models by examining the simpler models without using random coefficient for the auxiliary variable. Besides, the MCMC simulation for the Poisson-Lognormal models generally runs very slowly and requires much longer time to converge. Also, there has been studies indicating that Poisson-Gamma models are usually preferred because with the same moment structure, the Poisson-Gamma models are more conservative and result in minimax estimator of $\boldsymbol{\lambda}$ while the Poisson-Lognormal models do not (see Christiansen and Morris (1997)).

The p-values based on all measures do not show any difference between the models within each model structure (Poisson-Gamma vs. Poisson-Lognormal). The p-values for D_1 (the first level χ^2-discrepancy) are equal to 0 for all models, which suggests that the actual data has some extra dispersion that is not explained by the Poisson model at the school level. The p-value for D_5 (the negative minimum) are all equal to 0 too, which indicates the model is not adequate in describing the extremely small responses. The small p-values for D_{10} (the maximum of school ranges) also indicate the model fails to capture the extreme dispersion that is observed in the data. So measures D_1, D_5 and D_{10} overall indicate an inadequacy of the Poisson model at the school level. Accordingly we could consider a more complex model structure such as a zero-inflated Poisson model or a mixture of Poisson models for capturing

the distribution pattern of students within schools. Results for the Poisson-Lognormal models are the mostly similar. The two differences are that the p-values based on measures D_2 and D_3 (the second level and the overall χ^2-discrepancy) under the Poisson-Lognormal models are consistently larger than those under Poisson-Gamma models and the p-values based on measure D_6 (the range) under the Poisson-Lognormal models are consistently smaller than those under Poisson-Gamma models. These differences suggest distinct performance of the two model structures.

The Poisson-Lognormal models produce significantly higher second level and overall χ^2-discrepancy measures (D_2 and D_3) in the posterior predictions than for the realized data. The Poisson-Gamma models produce posterior predictive discrepancies spread fairly evenly around the corresponding 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 actual data. The underlying reason for this is the *higher skewness* for the Poisson-Gamma model toward zero better captures the very small rates of taking EP courses in some schools.

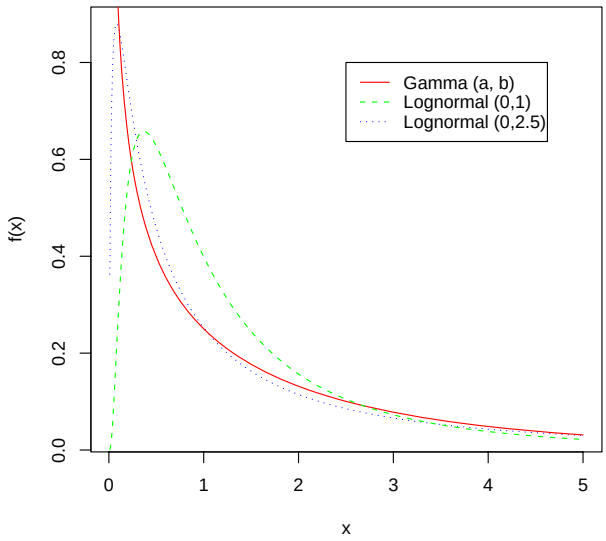


Figure 5.4 The probability density function of Gamma ($a = \frac{1}{e-1}, b = \frac{1}{(e-1)e^{1/2}}$), Lognormal$(0, 1)$ and Lognormal$(0, 2.5)$ distributions.

Figure 5.4 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 Lognormal$(0, 1)$ distribution matches the Gamma distribution better 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. In addition, the plot of the density functions shows that the Gamma and Lognormal distributions have very similar right tail behaviors but very distinct left tail behaviors. The shape and extension of the right tail for the two distributions are almost the same. But the Gamma distribution has much bigger probability for generating small values around zero. This characteristic of the distributions could be the reason that the Poisson-Gamma models produce predicted data values with larger ranges (D_6) than the Poisson-Lognormal models.

Table 5.5 shows the posterior predictive p-values based on the benchmarked discrepancy measures $D_1^{BHB}(\boldsymbol{y}, \boldsymbol{\theta})$ and $D_2^{BHB}(\boldsymbol{y}, \boldsymbol{\theta})$. Among the Poisson-Lognormal models, the model with an auxiliary variable and no random effects and the models with size random effect with and without auxiliary variable all have $D_1^{BHB}(\boldsymbol{y}^{obs}, \boldsymbol{\theta})$ bigger than 1/3, which means a large proportion (over one third) of strata have the benchmarked HB estimates falling outside of the 95% normal approximate posterior credible intervals of the HB estimators. Among the Poisson-Gamma models, the model with size random effect and no auxiliary variable, the models that have the auxiliary variable with a fixed coefficient with and without size random effect, and the models that have the auxiliary variable with random coefficients for size levels with and without random size effects all have $D_1^{BHB}(\boldsymbol{y}^{obs}, \boldsymbol{\theta})$ bigger than 1/3. The pooled model that has no random effects and no auxiliary variable has $D_1^{BHB}(\boldsymbol{y}^{obs}, \boldsymbol{\theta})$ bigger than 1/4 which is a quite large proportion, too. The rest of the candidate models have $D_1^{BHB}(\boldsymbol{y}^{obs}, \boldsymbol{\theta})$ fairly close or equal to 0. However, the p-values based on $D_1^{BHB}(\boldsymbol{y}^{obs}, \boldsymbol{\theta})$ have shown no evidence of model inadequacy for all models.

When discrepancy $D_2^{BHB}(\boldsymbol{y}, \boldsymbol{\theta})$ is used, among the Poisson-Lognormal models, the model with only an auxiliary variable and no random effects has the biggest average inflation of the PMSE. The models including only size effect with and without the auxiliary variable and the model with no auxiliary variable and no random effects also have much larger realized discrepancy values than other models that include an AEA effect. Among the Poisson-Gamma models, the model with only an auxiliary variable and no random effects also has the biggest average inflation of PMSE. The models that have no covariate variable with and without size random effect, the model that has the covariate variable with a fixed coefficient and size random effect, and the models that have the auxiliary variable with random regression coefficient for size levels with and without size random effects also have much larger realized discrepancy values than other models that include AEA random effect or allowing random regression coefficient for AEA levels. To see whether the realized discrepancy is extreme or not under the assumed model, we compared the $p_{post,2}^{BHB}$ values. None of the Poisson-Lognormal models has an extreme $p_{post,2}^{BHB}$ value. Even the models that have quite large realized discrepancy measures have $p_{post,2}^{BHB}$ values no less than 0.25. So the posterior predictive checks based on the p-values for the discrepancy measure $D_2^{BHB}(\boldsymbol{y}, \boldsymbol{\theta})$ showed no evidence of model incompatibilities with the actual data for all Poisson-Lognormal models. For the Poisson-Gamma models, the $p_{post,2}^{BHB}$ values showed that there are relatively smaller chances to have more extreme discrepancy values in the replicated predictions than for the actual data under the models with significantly larger realized discrepancies except the simplest (pooled) model which has no auxiliary variable and no random effects. The relatively small $p_{post,2}^{BHB}$ values indicate the five models with significantly larger realized discrepancies are less compatible with the actual data than other Poisson-Gamma models in terms of producing HB mean estimates that differ from the reliable direct estimates in larger regions in a large degree.

For the realized discrepancy values in Table 5.5, there is a general tendency that larger model tends to have smaller realized discrepancy values. But this pattern breaks when the most parsimonious models (the pooled models) in each of the two model structures are considered. The reason is that these models capture very few aspects of the data characteristics and fit

the data so poorly that they produce very large posterior variances. Since our discrepancy measure is based on the inflation of the PMSE relative to the posterior variance, even though the model is very wrong and produces extremely biased HB estimates, the large degree of inflation of the PMSE due to the bias correction is masked when the posterior variance itself is very large. Therefore, the benchmarked discrepancy has more power in detecting model inadequacy when the model does not produce a terrible small precision estimate of the model-based estimator. When the model is too wrong and produces intolerably large variance estimate for the model-based estimator, the benchmarked discrepancy loses some power for revealing the model failure.

For models that have the same basic model structure and fit the data well we usually choose the most succinct model for further analysis. By only referring to the $p^{BHB}_{post,2}$ values, we would end up with selecting the simplest (pooled) model in each model structure. These models produce estimates of strata means that are extremely different from the direct estimates and have very large variance estimates as well. However, if we preclude the simplest models which fit the data very badly and produce poor estimates with low precision for the quantity of interest, by only referring to the $p^{BHB}_{post,2}$ values, we would choose the Poisson-Gamma model with AEA random effects and no auxiliary variable or a Poisson-Lognormal model with an auxiliary variable and no random effects.

To compare these two models further, we computed $D_2^{BHB}(\boldsymbol{y}^{pred}, \boldsymbol{\theta}) - D_2^{BHB}(\boldsymbol{y}^{obs}, \boldsymbol{\theta})$ values for the two models. Figure 5.5 shows the sorted values of $D_2^{BHB}(\boldsymbol{y}^{pred}, \boldsymbol{\theta}) - D_2^{BHB}(\boldsymbol{y}^{obs}, \boldsymbol{\theta})$ under the Poisson-Gamma model with only AEA effect and no covariate variable and under the Poisson-Lognormal model with only an auxiliary variable and no random effects involved. The posterior predictive discrepancies under the Poisson-Lognormal model have a much wider spread than those under the Poisson-Gamma model. The pairwise comparison between other bigger Poisson-Lognormal models and the Poisson-Gamma model with AEA effect shows the same pattern. 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

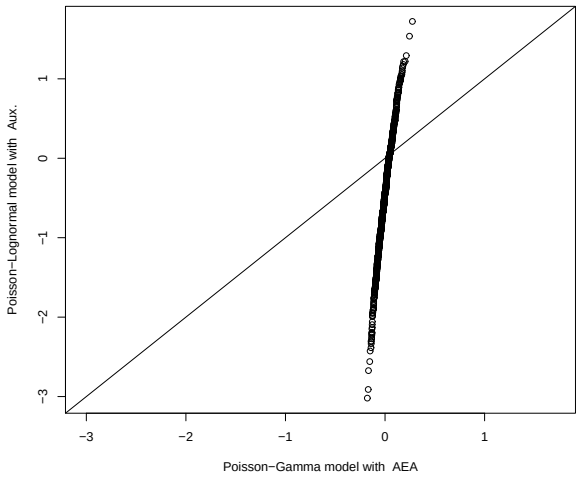


Figure 5.5 The scatter plot of sorted $(D_2^{BHB}(\boldsymbol{y}^{pred}, \boldsymbol{\theta}) - D_2^{BHB}(\boldsymbol{y}^{obs}, \boldsymbol{\theta}))$ values under Poisson-Gamma model with only AEA effect and no covariate variable and Poisson-Lognormal model with only an auxiliary variable and no random effects involved.

explains why the Poisson-Lognormal models with very large realized discrepancy values do not have extreme $p_{post,2}^{BHB}$ values.

By using the plot of the sorted values of $D_2^{BHB}(\boldsymbol{y}^{pred}, \boldsymbol{\theta}) - D_2^{BHB}(\boldsymbol{y}^{obs}, \boldsymbol{\theta})$ we made further comparison among the remaining Poisson-Gamma models that did not show any model inadequacy by examining the p-values in Table 5.5. We find that adding the covariate variable with a fixed coefficient and/or adding the size random effects (or adding the interactions as well) to the Poisson-Gamma model with only AEA random effects does not significantly improve the model fitness. Allowing a random coefficient for size levels for the covariate variable produces $D_2^{BHB}(\boldsymbol{y}^{pred}, \boldsymbol{\theta}) - D_2^{BHB}(\boldsymbol{y}^{obs}, \boldsymbol{\theta})$ values slightly more concentrated around zero, which indicates the data from the random coefficient model are a little less variable than the data from the fixed coefficient model and the no auxiliary model with AEA random effect.

However, the scatterplot of the antisymmetric quantity $D_2^{BHB}(\boldsymbol{y}^{pred}, \boldsymbol{\theta}) - D_2^{BHB}(\boldsymbol{y}^{obs}, \boldsymbol{\theta})$ can only show the relative behavior of the predictive discrepancy to the realized discrepancy but not the magnitude of the discrepancy values under the two models. Now imagine one model has realized discrepancy $D_2^{BHB}(\boldsymbol{y}^{obs}, \boldsymbol{\theta})$ at .3 and the predictive discrepancy values

center around the realized discrepancy and spread from .1 to .5. Imagine further that the second model has realized discrepancy at .7 and the predictive discrepancy values closely concentrate around .7 and range from .6 to .8. Then by comparing the sorted values of $D_2^{BHB}(\boldsymbol{y}^{pred}, \boldsymbol{\theta}) - D_2^{BHB}(\boldsymbol{y}^{obs}, \boldsymbol{\theta})$, we will only observe that the second model produce replicate data that are less variable but not be able to find that the second model produces replicate data that have consistently larger discrepancy values than the first model. Therefore, the plot of the sorted values of the antisymmetric quantity $D_2^{BHB}(\boldsymbol{y}^{pred}, \boldsymbol{\theta}) - D_2^{BHB}(\boldsymbol{y}^{obs}, \boldsymbol{\theta})$ could hide the relative magnitude of the discrepancy values under the models to be compared. Consequently, it could lose very important information concerning the degree of discrepancy between the data and the model assumptions.

To avoid such loss of information, we examined the overlaid histograms of the predictive discrepancy values with a vertical line representing the corresponding realized discrepancy under each of the models to be compared. Figure 5.6 shows the overlaid histograms for pairwise comparison of some of the Poisson-Gamma models. Adding an auxiliary variable with a fixed coefficient does not improve the model with only AEA random effect much. However, the model with a random coefficient for size levels produces consistently smaller and less variable discrepancy values than the model with only AEA effect, which provides "evidence" of in favor of the random coefficient model. Instead of adding the log-enrollment size as an auxiliary, including the size random effects also significantly improves the model with only AEA random effects by producing consistently smaller and less variable discrepancy values.

Further, the comparison between the model with size and AEA random effects and the model with AEA effect and the auxiliary variable with random coefficient for size levels are interesting. The realized discrepancy values under the two models are close. The discrepancy values generated under both models are fairly evenly distributed around the corresponding realized discrepancy values. However, the discrepancy values from the random coefficient model is more dispersed than those from the no auxiliary variable model. The latter is more concentrated, which suggests the latter is less variable, hence is preferred. Adding the auxiliary variable with either a fixed coefficient or a random coefficient for size or AEA levels would not

improve the model further. Therefore, we prefer the Poisson-Gamma model with both size and AEA random effects and no auxiliary variable, which fits the data well with a relatively simple model structure and produces reliable posterior estimates of means for small areas.

In conclusion, 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 from the candidate methods. The posterior predictive p-values using different discrepancy measures show advantages of detecting incompatibilities between the data and the 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 the PMSE due to benchmarking the HB estimator did much better in assessing overall fit of the models than other examined discrepancies. In addition, the graphical exploration has been proved very helpful in examining the performance of the models and shows an advantage over simple calculation of a p-value.

5.5 Concluding Remarks

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. Effective model selection is crucial in HB analysis. An HB posterior predictive model comparison method utilizing benchmarking was developed and shown to have the power to choose appropriate models in both an illustration and a real data analysis. The proposed method, which measures the inflation of the PMSE due to benchmarking the HB estimator, was compared to Bayesian model comparison based on the posterior predictive p-values using multiple other discrepancy measures, the L-criterion, and the deviance information criterion. The last two methods did a reasonable job in the illustrative example but showed less ability to detect inadequate models in analyzing the real data. The posterior predictive p-value using multi-

ple discrepancy measures especially the newly developed benchmarking discrepancy showed advantages in comparing models which are indistinguishable using the other two methods. Posterior predictive checking using a method of graphical examination yielded added insight into model comparison. The proposed benchmarking discrepancy is more effective in detecting mis-specified models with large bias and acceptable precision. The method could lose some power in revealing very bad models with intolerably large variance estimates. The method described here should be applicable beyond the survey and education contexts.

Table 5.4 The posterior predictive p-values for candidate models based on nine discrepancies described in Table 5.3 applied to the Iowa survey data.

Models	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9	D_{10}
Poisson-Lognormal models										
$\log(\lambda_{ijk}) = \beta_0 + \beta_1 x_{ijk} + v_{ijk}$	.000	1.000	1.000	1.000	.000	.377	.464	.645	.502	.000
$\log(\lambda_{ijk}) = \beta_0 + \beta_1 x_{ijk} + \tau_i + v_{ijk}$	.000	1.000	1.000	1.000	.000	.359	.493	.672	.524	.000
$\log(\lambda_{ijk}) = \beta_0 + \beta_1 x_{ijk} + \eta_j + v_{ijk}$	.000	1.000	1.000	1.000	.000	.369	.451	.675	.497	.000
$\log(\lambda_{ijk}) =$ $\beta_0 + \beta_1 x_{ijk} + \tau_i + \eta_j + v_{ijk}$	.000	1.000	1.000	1.000	.000	.368	.472	.648	.510	.000
$\log(\lambda_{ijk}) =$ $\beta_0 + \beta_1 x_{ijk} + \tau_i + \eta_j + \zeta_{ij} + v_{ijk}$	.000	1.000	1.000	1.000	.000	.359	.480	.661	.523	.000
$\log(\lambda_{ijk}) = \beta_0 + v_{ijk}$	.000	1.000	1.000	1.000	.000	.385	.499	.697	.545	.000
$\log(\lambda_{ijk}) = \beta_0 + \tau_i + v_{ijk}$	.000	1.000	1.000	1.000	.000	.346	.467	.657	.502	.000
$\log(\lambda_{ijk}) = \beta_0 + \eta_j + v_{ijk}$	.000	1.000	1.000	1.000	.000	.350	.466	.665	.501	.000
$\log(\lambda_{ijk}) = \beta_0 + \tau_i + \eta_j + v_{ijk}$	.000	1.000	1.000	1.000	.000	.375	.429	.649	.480	.000
$\log(\lambda_{ijk}) =$ $\beta_0 + \tau_i + \eta_j + \zeta_{ij} + v_{ijk}$	.000	1.000	1.000	1.000	.000	.380	.469	.637	.505	.000
Poisson-Gamma models										
$\log(\gamma_{ijk}) = \beta_0 + \beta_1 x_{ijk}$	.000	.489	.490	1.000	.000	1.000	.450	.679	.488	.000
$\log(\gamma_{ijk}) = \beta_0 + \beta_1 x_{ijk} + \tau_i$	.000	.487	.489	1.000	.000	1.000	.446	.685	.496	.000
$\log(\gamma_{ijk}) = \beta_0 + \beta_1 x_{ijk} + \eta_j$	.000	.461	.462	1.000	.000	1.000	.424	.665	.460	.000
$\log(\gamma_{ijk}) = \beta_0 + \beta_1 x_{ijk} + \tau_i + \eta_j$	.000	.456	.457	1.000	.000	1.000	.447	.665	.490	.000
$\log(\gamma_{ijk}) =$ $\beta_0 + \beta_1 x_{ijk} + \tau_i + \eta_j + \zeta_{ij}$	.000	.476	.472	1.000	.000	1.000	.444	.641	.478	.000
$\log(\gamma_{ijk}) = \beta_0$	.000	.537	.538	1.000	.000	1.000	.494	.706	.544	.000
$\log(\gamma_{ijk}) = \beta_0 + \tau_i$	.000	.502	.502	1.000	.000	1.000	.467	.684	.512	.000
$\log(\gamma_{ijk}) = \beta_0 + \eta_j$	.000	.486	.480	1.000	.000	1.000	.466	.681	.511	.000
$\log(\gamma_{ijk}) = \beta_0 + \tau_i + \eta_j$	.000	.457	.453	1.000	.000	1.000	.451	.650	.496	.000
$\log(\gamma_{ijk}) = \beta_0 + \tau_i + \eta_j + \zeta_{ij}$	.000	.480	.476	1.000	.000	1.000	.444	.630	.485	.000
$\log(\gamma_{ijk}) = \beta_0 + \beta_{1,i} x_{ijk}$	.000	.501	.500	1.000	.000	1.000	.458	.678	.499	.000
$\log(\gamma_{ijk}) = \beta_0 + \beta_{1,i} x_{ijk} + \tau_i$	.000	.524	.524	1.000	.000	1.000	.470	.687	.519	.000
$\log(\gamma_{ijk}) = \beta_0 + \beta_{1,i} x_{ijk} + \eta_j$	.000	.457	.454	1.000	.000	1.000	.424	.664	.463	.000
$\log(\gamma_{ijk}) = \beta_0 + \beta_{1,i} x_{ijk} + \tau_i + \eta_j$	.000	.454	.451	1.000	.000	1.000	.442	.658	.483	.000
$\log(\gamma_{ijk}) =$ $\beta_0 + \beta_{1,j} x_{ijk} + \tau_i + \eta_j + \zeta_{ij}$	.000	.476	.472	1.000	.000	1.000	.444	.641	.478	.000
$\log(\gamma_{ijk}) = \beta_0 + \beta_{1,j} x_{ijk}$	.000	.456	.453	1.000	.000	1.000	.447	.689	.494	.000
$\log(\gamma_{ijk}) = \beta_0 + \beta_{1,j} x_{ijk} + \tau_i$	.000	.423	.429	1.000	.000	1.000	.451	.674	.498	.000
$\log(\gamma_{ijk}) = \beta_0 + \beta_{1,j} x_{ijk} + \eta_j$	.000	.471	.469	1.000	.000	1.000	.443	.675	.482	.000
$\log(\gamma_{ijk}) = \beta_0 + \beta_{1,j} x_{ijk} + \tau_i + \eta_j$	.000	.468	.464	1.000	.000	1.000	.456	.686	.504	.000
$\log(\gamma_{ijk}) =$ $\beta_0 + \beta_{1,j} x_{ijk} + \tau_i + \eta_j + \zeta_{ij}$	.000	.460	.455	1.000	.000	1.000	.457	.651	.497	.000

Table 5.5 The posterior predictive p-values $p^{BHB}_{post,1}$ and $p^{BHB}_{post,2}$ based on the discrepancies $D^{BHB}_1(\boldsymbol{y}^{obs}, \boldsymbol{\theta})$ and $D^{BHB}_2(\boldsymbol{y}^{obs}, \boldsymbol{\theta})$ and the realized discrepancies $D^{BHB}_1(\boldsymbol{y}^{obs}, \boldsymbol{\theta})$ and $D^{BHB}_2(\boldsymbol{y}^{obs}, \boldsymbol{\theta})$ 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	$p^{BHB}_{post,1}$	$D^{BHB}_1(\boldsymbol{y}^{obs}, \boldsymbol{\theta})$	$p^{BHB}_{post,2}$	$D^{BHB}_2(\boldsymbol{y}^{obs}, \boldsymbol{\theta})$
Poisson-Lognormal Models				
$\log(\lambda_{ijk}) = \beta_0 + \beta_1 x_{ijk} + v_{ijk}$	.578	**.423**	.265	**5.990**
$\log(\lambda_{ijk}) = \beta_0 + \beta_1 x_{ijk} + \tau_i + v_{ijk}$	.721	**.385**	.649	**3.850**
$\log(\lambda_{ijk}) = \beta_0 + \beta_1 x_{ijk} + \eta_j + v_{ijk}$	.822	.038	.879	.733
$\log(\lambda_{ijk}) = \beta_0 + \beta_1 x_{ijk} + \tau_i + \eta_j + v_{ijk}$	1.000	.000	.639	.474
$\log(\lambda_{ijk}) = \beta_0 + \beta_1 x_{ijk} + \tau_i + \eta_j + \zeta_{ij} + v_{ijk}$	1.000	.000	.994	.099
$\log(\lambda_{ijk}) = \beta_0 + v_{ijk}$	.665	.154	.561	**2.124**
$\log(\lambda_{ijk}) = \beta_0 + \tau_i + v_{ijk}$	.456	**.385**	.315	**3.255**
$\log(\lambda_{ijk}) = \beta_0 + \eta_j + v_{ijk}$	.836	.038	.762	.987
$\log(\lambda_{ijk}) = \beta_0 + \tau_i + \eta_j + v_{ijk}$	1.000	.000	.800	.311
$\log(\lambda_{ijk}) = \beta_0 + \tau_i + \eta_j + \zeta_{ij} + v_{ijk}$	1.000	.000	.952	.152
Poisson-Gamma Models				
$\log(\gamma_{ijk}) = \beta_0 + \beta_1 x_{ijk}$	.947	**.423**	**.033**	**7.876**
$\log(\gamma_{ijk}) = \beta_0 + \beta_1 x_{ijk} + \tau_i$	.623	**.423**	**.050**	**4.735**
$\log(\gamma_{ijk}) = \beta_0 + \beta_1 x_{ijk} + \eta_j$	.337	.038	.261	.825
$\log(\gamma_{ijk}) = \beta_0 + \beta_1 x_{ijk} + \tau_i + \eta_j$	1.000	.000	.318	.376
$\log(\gamma_{ijk}) = \beta_0 + \beta_1 x_{ijk} + \tau_i + \eta_j + \zeta_{ij}$	1.000	.000	.631	.229
$\log(\gamma_{ijk}) = \beta_0$	.648	**.269**	.408	**2.814**
$\log(\gamma_{ijk}) = \beta_0 + \tau_i$	.654	**.385**	.115	**3.941**
$\log(\gamma_{ijk}) = \beta_0 + \eta_j$	1.000	.000	.584	.802
$\log(\gamma_{ijk}) = \beta_0 + \tau_i + \eta_j$	1.000	.000	.422	.295
$\log(\gamma_{ijk}) = \beta_0 + \tau_i + \eta_j + \zeta_{ij}$	1.000	.000	.387	.217
$\log(\gamma_{ijk}) = \beta_0 + \beta_{1,i} x_{ijk}$	.805	**.385**	**.054**	**4.028**
$\log(\gamma_{ijk}) = \beta_0 + \beta_{1,i} x_{ijk} + \tau_i$	.772	**.385**	**.036**	**4.100**
$\log(\gamma_{ijk}) = \beta_0 + \beta_{1,i} x_{ijk} + \eta_j$	1.000	.000	.374	.317
$\log(\gamma_{ijk}) = \beta_0 + \beta_{1,i} x_{ijk} + \tau_i + \eta_j$	1.000	.000	.444	.314
$\log(\gamma_{ijk}) = \beta_0 + \beta_{1,i} x_{ijk} + \tau_i + \eta_j + \zeta_{ij}$	1.000	.000	.321	.229
$\log(\gamma_{ijk}) = \beta_0 + \beta_{1,j} x_{ijk}$	1.000	.000	.394	.836
$\log(\gamma_{ijk}) = \beta_0 + \beta_{1,j} x_{ijk} + \tau_i$	1.000	.000	.261	.480
$\log(\gamma_{ijk}) = \beta_0 + \beta_{1,j} x_{ijk} + \eta_j$	1.000	.000	.521	.684
$\log(\gamma_{ijk}) = \beta_0 + \beta_{1,j} x_{ijk} + \tau_i + \eta_j$	1.000	.000	.445	.308
$\log(\gamma_{ijk}) = \beta_0 + \beta_{1,j} x_{ijk} + \tau_i + \eta_j + \zeta_{ij}$	1.000	.000	.656	.201

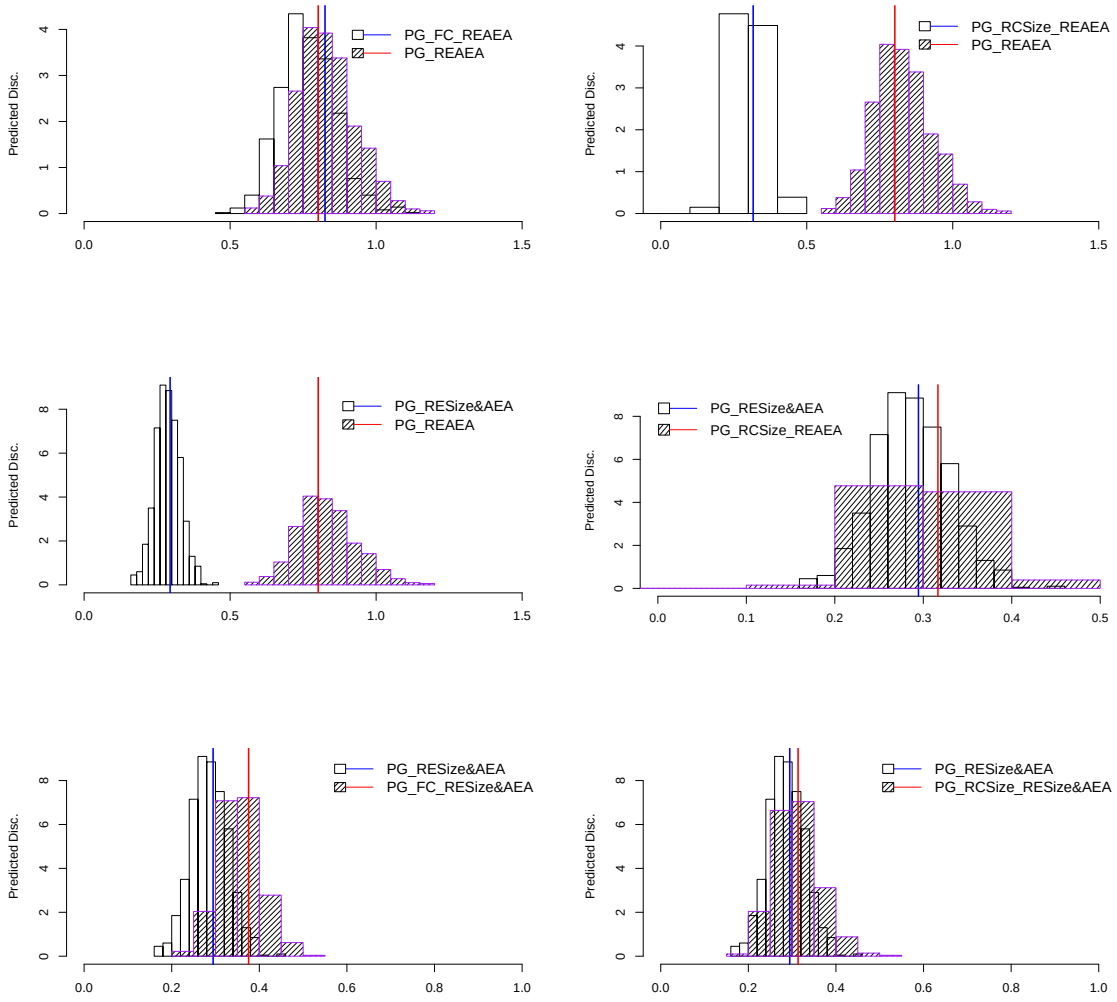


Figure 5.6 The overlaid histograms of predictive discrepancy values with vertical lines representing the realized discrepancies for Poisson-Gamma models. In the legends, "PG", "FC", "RC", and "RE" stand for "Poisson-Gamma" model, "fixed coefficient", "random coefficient", and "random effect", respectively. For example, "PG_REAEA" represents the Poisson-Gamma model with only AEA random effect and no auxiliary variable; "PG_RCSize_REAEA" denotes the Poisson-Gamma model with AEA random effect and an auxiliary variable with random coefficient for size levels.

CHAPTER 6. Conclusions and Future Study

6.1 Conclusions

In this thesis we studied three primary topics. First, we considered estimators, design issues, and variance estimation in complex survey designs with a single primary sampling unit in some strata. Second, we examined small area estimation using hierarchical models and Bayesian analysis. Third, we proposed and studied a method for model selection based on posterior predictive checks using a benchmarking discrepancy measure. A survey that collected data from the transcripts of Iowas public high school students served as a data source and motivation for our study.

To produce reasonable variance estimates for individual strata with one PSU sampled, in Chapter 3, we began with the commonly used collapsed strata estimator and proposed a synthetic redistribution of the variance of the collapsed group to the individual strata within the group. This was referred to as the collapsed strata synthetic variance (CSSV) method. The collapsed strata estimator usually overestimates the variance of the collapsed group when the strata in the group were not quite "similar" in terms of the characteristics being measured and the size of the stratum. The method used for redistribution is not unique and should be tailored to the specific design and data structure.

We then considered an alternative method that uses a generalized variance function (GVF) to explicitly model the relationship between the variance and the expectation of the total estimator. The relationship is estimated using a large group of strata sharing similar characteristics based on the direct estimates for some members of the group with more than one PSU in the sample. The estimated relationship is used to predict the variances of strata with one PSU sampled based on their estimated totals. Since the designs in strata used for fitting the model

and for predicting the variance were different in terms of the sample sizes of PSUs, we proposed to either make adjustment to the coefficients of the fitted model for appropriate prediction or collapse the one-PSU strata into groups such that the collapsed group has exactly the same "design" (sample PSUs) as the ones that were used to fit the model. In the latter case the prediction of variance for the collapsed group is followed by synthetic variance redistribution. A restricted GVF (RGVF) was used to avoid negative predictions of variance, which could cause a serious problem in utilizing the traditional GVF without restrictions. This situation could be even worse when the collapsing procedure was employed for predicting variance of a group of strata. The proposed adjusted RGVF methods were compared with direct estimation and then CSSV method in simulation studies and the ISBE application. The modified RGVF methods were suggested because they showed advantages in producing a consistently higher coverage rate of confidence intervals and more stable estimates of variance. The CSSV method was shown to very frequently overestimate the variance in the simulation. The collapsing strata RGVF synthetic variance (CRGVFSV) estimation seemed to be less conservative than the adjusted (coefficients) RGVF method in our limited simulation. To claim the superiority of one over the other, further examination is needed.

Since the GVF method is dependent on the direct estimates of totals and variances and is subject to the shortcomings of the design-based estimator in substance, we considered using an explicit model-based method to produce more reliable estimates and variances of the estimates for small area statistics. We chose to use hierarchical Bayes methods because it is straightforward to implement them and they provide a unified framework for large and small sample cases. We began with two families of generalized linear mixed models (GLMMs) to characterize the correlation structure of the ISBE survey data. The hierarchical Bayesian (HB) approach was used to estimate the stratum means and variances for the HB estimates. We then illustrated the precision of the HB estimator relative to the ratio estimator using a simulated finite population that has a similar structure as the ISBE's. In Chapter 4, we did an exploratory data analysis for the actual ISBE survey data using models from the proposed two families of GLMMs. An informal model building process was illustrated to select most satis-

factory models from a class of plausible models. The analysis results showed strong evidence of variability along both size and AEA factors, which were both design factors in the survey.

In Chapter 5, we focused on the study of effective model selection within the Bayesian framework. A hierarchical Bayesian posterior predictive model comparison method utilizing benchmarking in proposing a discrepancy measure was developed. Using a method of graphical examination in posterior predictive checking was also suggested for adding insight into model comparison. The proposed method was shown to have the power of choosing appropriate models in both an illustrative example and the analysis of the actual ISBE survey data. In the latter case, administrative variables from the Common Core Data (CCD) from the National Center for Education Statistics (NCES) were used as covariates in the models. Traditional methods such as posterior predictive p-values using many other discrepancy measures, the L-criterion, and the deviance information criterion all failed to detect the model differences. As a side note, the proposed discrepancy that measures the inflation of posterior mean squared error due to benchmarking relative to the posterior variance is more effective in detecting model inadequacy when the model is misspecified and produces biased results with acceptable precision, in which circumstances one is more likely to draw misleading conclusions. The method could lose some power for revealing model failure when the model is actually too wrong and produces intolerably large variance estimates for the HB estimator. However, these cases are often very easy to detect and very unlikely to cause any real trouble in the course of the data analysis. The ISBE survey design was complex in terms of containing stratification and clustering, which are common design elements for large scale surveys and many applications in other fields. The hierarchical Bayesian estimation and model selection methods discussed in Chapters 4 and 5 should be able to be adapted to many other surveys or applications whose data have stratification/blocking and clustering structures.

6.2 Future Study

Future study will focus on extending the hierarchical Bayesian small area estimation methods. In the work of hierarchical Bayesian posterior predictive checking utilizing a benchmarking

discrepancy, we examined the performance of the proposed method in an illustrative example and a real data analysis. These were all done based on a single sample from a finite population. To make it a compelling case, an evaluation of the proposed method based on a large number of samples, say 200, from a simulated population can be done. The proportion of cases where the proposed method can successfully identify the true model can be reported. We can further study the properties of the posterior predictive p-values based on the proposed benchmarking discrepancy through simulation. For example, we can study the sampling distribution of the p-values under the null (true) model based on a large number of samples in different scenarios of small, moderate and large sample sizes. An approximate uniform distribution would indicate nice properties. We can also compute the power of the the p-values under some false models with different choices of nominal values in scenarios of different sample sizes. To conduct such simulation studies by implementing complex GLMMs through MCMC would require a very heavy load of computation and could be very time consuming. We can consider starting from simpler model structures such as two-level normal random effect models.

As we have discussed in Section 5.1.1, the posterior predictive p-values suffer from the drawbacks of double use of data. They could be very conservative in detecting model failure. In Larsen and Lu (2007), we recommended cross-validated posterior predictive model selection. The method computes the p-values for individual groups of data using the posterior predictive distributions obtained by leaving out groups of data, thereby avoiding the double use of data. Extreme p-values for any groups of data could be an indication of model inadequacy. However, the implementation of the method requires refitting the model without each group of data, which is time consuming when the number of groups is large. Methods such as *importance weighting* and *importance sampling* (Stern and Cressie (2000)) can be used to approximate the posterior distribution that would be obtained if the analysis were repeated while leaving out the group without actually refitting the model without the group. The properties of the cross-validated posterior predictive p-values based on the benchmarking discrepancy can be examined through simulation studies analogous to those above for the posterior predictive p-values. The issue of using multiple test statistics concerning individual groups should be

addressed. Adjustments for multiple testing will affect the power for comparing models.

In our study, we have examined the selection of model structures and variables. But we have not touched the issue of choosing transformations of predictive variables. The Box-Cox transformation or the simpler power transformation can be used. The comparison can be made on a discrete grid of power parameters. However, the computation could be heavy especially if we are choosing variables as well as transformations of the variables. One possibility is to consider the power parameters as model parameters, assign them some appropriate prior distributions and fit the model through MCMC to obtain the posterior estimates of the power parameters. However, to implement this process, we need to develop an MCMC sampling algorithm that can jump among the power parameters with acceptable efficiency. The development of an efficient strategy to combine the selection of variables and transformations is also of interest.

In large-scale studies, there are many variables involved that potentially can be used for modeling. It is practically inefficient or impossible to compare all possible models with various combinations of variables. We hope to explore methods that can narrow the range of candidate models, which would allow us to do one-by-one model comparison using the proposed methods. There have been some contributions to the field, such as the *stochastic search variable selection* (SSVS) proposed by George and McCulloch (1993). The idea was to use latent variables to identify choices of subsets of predictors and select the promising subsets with higher posterior probability. An initial selection of candidate models could potentially help SSVS overcome some difficulties with multicollinearity of predictors. 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. Instead of choosing one single model and assuming the model is true to carry on the analysis, we could account for model uncertainty by averaging the predicted values over a group of promising models weighted by the posterior probabilities of the models.

APPENDIX Derivations for the Posterior Mean and Variance-Covariance of School Level Poisson Rate $\boldsymbol{\lambda}$ for Hierarchical Bayesian Models

For the Poisson-Lognormal model defined in (4.1) and (4.2) in Section 4.2, the posterior mean of school rate λ_{ijk} is

$$\begin{aligned} E\left(\lambda_{ijk}|\boldsymbol{y}^{obs}\right) &= E\{E(\lambda_{ijk}|\boldsymbol{\beta},\tau_i,\eta_j,\zeta_{ij},\sigma_v^2,\boldsymbol{y}^{obs})|\boldsymbol{y}^{obs}\} \\ &= E\{\exp(\boldsymbol{\beta}'\boldsymbol{x}_{ijk}+\tau_i+\eta_j+\zeta_{ij}+\frac{1}{2}\sigma_v^2)|\boldsymbol{y}^{obs}\}, \end{aligned} \tag{A.1}$$

where the expectations in (A.1) and the following (A.2) and (A.3) are taken with respect to the joint posterior distribution of $\boldsymbol{\beta}$, $\boldsymbol{\tau}$, $\boldsymbol{\eta}$, $\boldsymbol{\zeta}$ and σ_v^2 given the data. The posterior variance of λ_{ijk} is

$$\begin{aligned} & V\left(\lambda_{ijk}|\boldsymbol{y}^{obs}\right) \\ = \; & V\{E(\lambda_{ijk}|\boldsymbol{\beta},\tau_i,\eta_j,\zeta_{ij},\sigma_v^2,\boldsymbol{y}^{obs})|\boldsymbol{y}^{obs}\}+E\{V(\lambda_{ijk}|\boldsymbol{\beta},\tau_i,\eta_j,\zeta_{ij},\sigma_v^2,\boldsymbol{y}^{obs})|\boldsymbol{y}^{obs}\} \\ = \; & V\{\exp(\boldsymbol{\beta}'\boldsymbol{x}_{ijk}+\tau_i+\eta_j+\zeta_{ij}+\frac{1}{2}\sigma_v^2)|\boldsymbol{y}^{obs}\}+E\{\exp[2(\boldsymbol{\beta}'\boldsymbol{x}_{ijk}+\tau_i+\eta_j+\zeta_{ij})+\sigma_v^2](e^{\sigma_v^2}-1)|\boldsymbol{y}^{obs}\} \\ = \; & E\{\exp[2(\boldsymbol{\beta}'\boldsymbol{x}_{ijk}+\tau_i+\eta_j+\zeta_{ij}+\sigma_v^2)]|\boldsymbol{y}^{obs}\}-[E\{\exp(\boldsymbol{\beta}'\boldsymbol{x}_{ijk}+\tau_i+\eta_j+\zeta_{ij}+\frac{1}{2}\sigma_v^2)|\boldsymbol{y}^{obs}\}]^2 \\ = \; & E\{\exp[2(\boldsymbol{\beta}'\boldsymbol{x}_{ijk}+\tau_i+\eta_j+\zeta_{ij}+\sigma_v^2)]|\boldsymbol{y}^{obs}\}-[E\left(\lambda_{ijk}|\boldsymbol{y}^{obs}\right)]^2. \end{aligned} \tag{A.2}$$

The posterior covariance between λ_{ijk} and $\lambda_{i'j'k'}$ is given by

$$
\begin{aligned}
& C\left(\lambda_{ijk}, \lambda_{i'j'k'} | \boldsymbol{y}^{obs}\right) \\
= \; & C\{E(\lambda_{ijk}|\boldsymbol{\beta}, \tau_i, \eta_j, \zeta_{ij}, \tau_{i'}, \eta_{j'}, \zeta_{i'j'}, \sigma_v^2, \boldsymbol{y}^{obs}), E(\lambda_{i'j'k'}|\boldsymbol{\beta}, \tau_i, \eta_j, \zeta_{ij}, \tau_{i'}, \eta_{j'}, \zeta_{i'j'}, \sigma_v^2, \boldsymbol{y}^{obs})|\boldsymbol{y}^{obs}\} \\
& + E\{C(\lambda_{ijk}, \lambda_{i'j'k'}|\boldsymbol{\beta}, \tau_i, \eta_j, \zeta_{ij}, \tau_{i'}, \eta_{j'}, \zeta_{i'j'}, \sigma_v^2, \boldsymbol{y}^{obs})|\boldsymbol{y}^{obs}\} \\
= \; & C\{\exp(\boldsymbol{\beta}'\boldsymbol{x}_{ijk} + \tau_i + \eta_j + \zeta_{ij} + \frac{1}{2}\sigma_v^2), \exp(\boldsymbol{\beta}'\boldsymbol{x}_{i'j'k'} + \tau_{i'} + \eta_{j'} + \zeta_{i'j'} + \frac{1}{2}\sigma_v^2)|\boldsymbol{y}^{obs}\} \\
= \; & E\{\exp[\boldsymbol{\beta}'(\boldsymbol{x}_{ijk} + \boldsymbol{x}_{i'j'k'}) + \tau_i + \tau_{i'} + \eta_j + \eta_{j'} + \zeta_{ij} + \zeta_{i'j'} + \sigma_v^2]|\boldsymbol{y}^{obs}\} \\
& - E\{\exp(\boldsymbol{\beta}'\boldsymbol{x}_{ijk} + \tau_i + \eta_j + \zeta_{ij} + \frac{1}{2}\sigma_v^2)|\boldsymbol{y}^{obs}\} \\
& \cdot E\{\exp(\boldsymbol{\beta}'\boldsymbol{x}_{i'j'k'} + \tau_{i'} + \eta_{j'} + \zeta_{i'j'} + \frac{1}{2}\sigma_v^2)|\boldsymbol{y}^{obs}\} \qquad \text{(A.3)} \\
= \; & E\{\exp[\boldsymbol{\beta}'(\boldsymbol{x}_{ijk} + \boldsymbol{x}_{i'j'k'}) + \tau_i + \tau_{i'} + \eta_j + \eta_{j'} + \zeta_{ij} + \zeta_{i'j'} + \sigma_v^2]|\boldsymbol{y}^{obs}\} \\
& - E\left(\lambda_{ijk}|\boldsymbol{y}^{obs}\right) E\left(\lambda_{i'j'k'}|\boldsymbol{y}^{obs}\right).
\end{aligned}
$$

For the Poisson-Gamma model defined in (4.1) and (4.3) in Section 4.2, the posterior mean of school rate λ_{ijk} is

$$
\begin{aligned}
E\left(\lambda_{ijk}|\boldsymbol{y}^{obs}\right) &= E\{E(\lambda_{ijk}|\gamma_{ijk}, \alpha, \boldsymbol{y}^{obs})|\boldsymbol{y}^{obs}\} \\
&= E\{E(\gamma_{ijk}|\boldsymbol{\beta}, \tau_i, \eta_j, \zeta_{ij}, \alpha, \boldsymbol{y}^{obs})|\boldsymbol{y}^{obs}\} \\
&= E\{\exp(\boldsymbol{\beta}'\boldsymbol{x}_{ijk} + \tau_i + \eta_j + \zeta_{ij}|\boldsymbol{y}^{obs}\}, \qquad \text{(A.4)}
\end{aligned}
$$

where the expectations in (A.4) and the following (A.5) and (A.6) are taken with respect to the joint posterior distribution of $\boldsymbol{\beta}$, $\boldsymbol{\tau}$, $\boldsymbol{\eta}$, $\boldsymbol{\zeta}$ and α given the data. The posterior variance of λ_{ijk} is

$$
\begin{aligned}
& V\left(\lambda_{ijk}|\boldsymbol{y}^{obs}\right) \\
= \; & V\{E(\lambda_{ijk}|\gamma_{ijk}, \alpha, \boldsymbol{y}^{obs})|\boldsymbol{y}^{obs}\} + E\{V(\lambda_{ijk}|\gamma_{ijk}, \alpha, \boldsymbol{y}^{obs})|\boldsymbol{y}^{obs}\} \\
= \; & E\{(1 + 1/\alpha)\gamma_{ijk}^2)|\boldsymbol{y}^{obs}\} - E^2(\gamma_{ijk}|\boldsymbol{y}^{obs}) \\
= \; & E\{(1 + 1/\alpha)\exp[2(\boldsymbol{\beta}'\boldsymbol{x}_{ijk} + \tau_i + \eta_j + \zeta_{ij})]|\boldsymbol{y}^{obs}\} \\
& - E^2\{\exp(\boldsymbol{\beta}'\boldsymbol{x}_{ijk} + \tau_i + \eta_j + \zeta_{ij})|\boldsymbol{y}^{obs}\} \qquad \text{(A.5)} \\
= \; & E\{(1 + 1/\alpha)\exp[2(\boldsymbol{\beta}'\boldsymbol{x}_{ijk} + \tau_i + \eta_j + \zeta_{ij})]|\boldsymbol{y}^{obs}\} - [E\left(\lambda_{ijk}|\boldsymbol{y}^{obs}\right)]^2.
\end{aligned}
$$

The posterior covariance between λ_{ijk} and $\lambda_{i'j'k'}$ is given by

$$
\begin{aligned}
& C\left(\lambda_{ijk}, \lambda_{i'j'k'} | \boldsymbol{y}^{obs}\right) \\
= \; & C\{E(\lambda_{ijk}|\gamma_{ijk}, \gamma_{i'j'k'}, \alpha, \boldsymbol{y}^{obs}), E(\lambda_{i'j'k'}|\gamma_{ijk}, \gamma_{i'j'k'}, \alpha, \boldsymbol{y}^{obs})|\boldsymbol{y}^{obs}\} \\
& + E\{C(\lambda_{ijk}, \lambda_{i'j'k'}|\gamma_{ijk}, \gamma_{i'j'k'}, \alpha, \boldsymbol{y}^{obs})|\boldsymbol{y}^{obs}\} \\
= \; & C\{\exp(\boldsymbol{\beta'}\boldsymbol{x}_{ijk} + \tau_i + \eta_j + \zeta_{ij}), \exp(\boldsymbol{\beta'}\boldsymbol{x}_{i'j'k'} + \tau_{i'} + \eta_{j'} + \zeta_{i'j'})|\boldsymbol{y}^{obs}\} \\
= \; & E\{\exp[\boldsymbol{\beta'}(\boldsymbol{x}_{ijk} + \boldsymbol{x}_{i'j'k'}) + \tau_i + \tau_{i'} + \eta_j + \eta_{j'} + \zeta_{ij} + \zeta_{i'j'}]|\boldsymbol{y}^{obs}\} \\
& - E\{\exp(\boldsymbol{\beta'}\boldsymbol{x}_{ijk} + \tau_i + \eta_j + \zeta_{ij}|\boldsymbol{y}^{obs}\} E\{\exp(\boldsymbol{\beta'}\boldsymbol{x}_{i'j'k'} + \tau_{i'} + \eta_{j'} + \zeta_{i'j'}|\boldsymbol{y}^{obs}\} \quad \text{(A.6)} \\
= \; & E\{\exp[\boldsymbol{\beta'}(\boldsymbol{x}_{ijk} + \boldsymbol{x}_{i'j'k'}) + \tau_i + \tau_{i'} + \eta_j + \eta_{j'} + \zeta_{ij} + \zeta_{i'j'}]|\boldsymbol{y}^{obs}\} \\
& - E\left(\lambda_{ijk}|\boldsymbol{y}^{obs}\right) E\left(\lambda_{i'j'k'}|\boldsymbol{y}^{obs}\right).
\end{aligned}
$$

BIBLIOGRAPHY

Bayarri, M. J. and Castellanos, M. E. (2007), "Bayesian Checking of the Second Levels of Hierarchical Models," *Statistical Science*, 22, 322–343.

Christiansen, C. L. and Morris, C. N. (1997), "Hierarchical Poisson Regression Modeling," *Journal of the American Statistical Association*, 92, 618–632.

Cochran, W. G. (1977), *Sampling Techniques*, New York: J. Wiley, 3rd ed.

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. (2004), "Exploratory data analysis for complex models," *Journal of Computational and Graphical Statistics*, 13, 755–779.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2004), *Bayesian Data Analysis*, London: Chapman and Hall, 2nd ed.

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

George, E. I. and McCulloch, R. E. (1993), "Variable Selection Via Gibbs Sampling," *Journal of the American Statistical Association*, 88, 881–889.

Ghosh, M. and Rao, J. (1994), "Small Area Estimation: An Appraisal," *Statistical Science*, 9, 55–93.

Guttman, I. (1967), "The Use of the Concept of a Future Observation in Goodness-of-Fit Problems," *Journal of the Royal Statistical Society. Series B (Methodological)*, 29, 83–100.

Hansen, M. H., Hurwitz, W. N., and Madow, W. G. (1953), *Sample survey methods and theory*, New York: Wiley, 1st ed.

Hartley, H. O., Rao, J. N. K., and Kiefer, G. (1969), "Variance Estimation with One Unit Per Stratum," *Journal of the American Statistical Association*, 64, 841–851.

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.

Horvitz, D. G. and Thompson, D. J. (1952), "A Generalization of Sampling Without Replacement From a Finite Universe," *Journal of the American Statistical Association*, 47, 663–685.

Isaki, C. T. (1983), "Variance Estimation Using Auxiliary Information," *Journal of the American Statistical Association*, 78, 117–123.

Isaki, C. T. and Fuller, W. A. (1982), "Survey Design Under the Regression Superpopulation Model," *Journal of the American Statistical Association*, 77, 89–96.

Jiang, J. and Lahiri, P. (2006), "Mixed Model Prediction and Small Area Estimation," *Test*, 15, 1–96.

Kass, R. E. and Raftery, A. E. (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90, 773–795.

Larsen, M. D. and Lu, L. (2007), "Comment: Bayesian Checking of the Second Levels of Hierarchical Models: Cross-validated posterior predictive checks using discrepancy measures," *Statistical Science*, 22, 359–362.

Laud, P. W. and Ibrahim, J. G. (1995), "Predictive Model Selection," *Journal of the Royal Statistical Society. Series B (Methodological)*, 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 Reserach Methods Section, American Statistical Association*.

— (2007), "Variance Estimation in a high school student survey with one-per-stratum strata," *Proceedings of the Third International Conference on Establishment Surveys (ICES-III)*.

Meng, X. L. (1994), "Posterior Predictive p-Values," *The Annals of Statistics*, 22, 1142–1160.

Rao, J. (2003), *Small Area Estimation*, Hoboken, NJ: Wiley, 1st ed.

Rubin, D. B. (1984), "Bayesianly Justifiable and Relevant Frequency Calculations for the Applies Statistician," *The Annals of Statistics*, 12, 1151–1172.

Särndal, C., Swensson, B. S., and Wretman, J. H. (1992), *Model Assisted Survey Sampling*, New York: Springer, 1st ed.

Särndal, C. E., Swensson, B. S., and Wretman, J. H. (1989), "The Weighted Residual Technique for Estimating the Variance of the General Regression Estimator of the Finite Population Total," *Biometrika*, 76, 527–537.

Shapiro, G. M., Singh, R. P., and Bateman, D. (1980), "Empirical research involving an alternative variance estimator to the collapsed stratum variance estimator," *Proceedings of the Survey Research Methods Section, American Statistical Association*, 793–798.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Linde, A. v. d. (2002), "Bayesian Measures of Model Complexity and Fit," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 64, 583–639.

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

Valliant, R. (1987), "Generalized Variance Functions in Stratified Two-Stage Sampling," *Journal of the American Statistical Association*, 82, 499–508.

Wolter, K. M. (1985), *Introduction to variance estimation*, Springer Series in Statistics, 1st ed.

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.

Zhao, O. Y., Schreuder, H. T., and Li, J. F. (1991), "Regression estimation under sampling with one unit per stratum," *Communications in Statistics - Theory and Methods*, 20, 2431–2449.

ACKNOWLEDGEMENTS

This work was supported in part by Iowa's State Board of Education and a dissertation 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.

I would like to express my sincere gratitude to my advisor Dr. Michael Larsen for his guidance, support, patience and encouragement throughout my graduate studies. He not only taught me statistical knowledge and inspired me with thoughts and ideas on my study but also gave me insights on conducting good academic research. He has always been there listening and giving advice on every little problem that I came across in my course work, research study, career planning, and life in general. Without his help and support, this thesis would not have been possible.

I would also like to thank Dr. Jean Opsomer, Dr. Taps Maiti, Dr. Fred Lorenz, Dr. Amy Froelich, and Dr. Craig Gundersen for serving on my committee and for the valuable comments and advice and the support at all levels that they have given in this process. I am also very thankful to my fellow graduate students for their encouragement.