ABSTRACT

Title of Document: HIERARCHICAL BAYES ESTIMATION AND EMPIRICAL BEST PREDICTION OF SMALL-AREA PROPORTIONS

Benmei Liu, Doctor of Philosophy, 2009

Directed By: Professor Partha Lahiri
Joint Program in Survey Methodology

Estimating proportions of units with a given characteristic for small areas using small area estimation (SAE) techniques is a common problem in survey research. The direct survey estimates, usually based on area-specific sample data, are very imprecise or even unavailable due to the small or zero sample sizes in the areas. In order to provide precise estimates, a variety of model-dependent techniques, using Bayesian and frequentist approaches, have been developed. Among those, empirical best prediction (EBP) and hierarchical Bayes (HB) methods relying on mixed models have been considered for estimating small area proportions.

Mixed models can be broadly classified as area or unit level models in SAE. When an area level model is used to produce estimates of proportions for small areas, it is commonly assumed that the survey weighted proportion for each sampled small area has a normal distribution and that the sampling variance of this proportion is known. However, these assumptions are problematic when the small area sample size

is small or when the true proportion is near 0 or 1. In addition, normality is commonly assumed for the random effects in area level and unit level mixed models. However, this assumption maybe violated for some cases.

To address those issues, in this dissertation, we first explore some alternatives to the well-known Fay-Herriot area level model. The aim is to consider models that are appropriate for survey-weighted proportions and can capture different sources of uncertainty, including the uncertainty that arises from the estimation of the sampling variances of the design-based estimators. Then we develop an adaptive HB method for SAE using data from a simple stratified design. The main goal is to relax the usual normality assumption for the random effects and instead determine the distribution of the random effects adaptively from the survey data. The Jiang-Lahiri type frequentist's alternative to the hierarchical Bayesian methods is also developed. Finally we propose a generalized linear mixed model that is suitable for binary data collected from a two-stage sampling design.

HIERARCHICAL BAYES ESTIMATION AND EMPIRICAL BEST PREDICTION OF SMALL-AREA PROPORTIONS

By

Benmei Liu

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2009

Advisory Committee:
Professor Partha Lahiri, Chair
Research Professor Graham Kalton
Research Professor Keith Rust
Professor Paul Smith
Professor Richard Valliant

Dedication

This dissertation is dedicated to my parents, Yuangui Liu and Xuying Chen, and to my husband, Tianbing Qi.

Acknowledgements

First and foremost, I must express my sincerest gratitude to my academic advisor, Professor Partha Lahiri. Throughout years, he has continuously guided me in my study and research through the use of his profound knowledge and insightful thinking. I really appreciate the enormous amount of time and patience he has spent on me, without which I would never dream of achieving the current stage in my career.

I also want to extend my gratitude to my other committee members: Dr. Graham Kalton, Dr. Keith Rust, Dr. Paul Smith, and Dr. Richard Valliant for their constructive comments and helps during the whole process of my dissertation research.

My special thanks go to Professor Roger Tourangeau for his consistent encouragement and support through my Ph.D. study.

For David Morganstein and my senior colleagues at Westat, I want to thank them for coordinating me into a flexibility that allowed me to manage my work and study simultaneously during the primary stage of my Ph.D. study.

I shall also thank the members in our JPSM family: Rupa Jethwa Eapen and Sarah Gebremicael offered me their administrative help; Duane Gilbert provided me all the IT hardware and software that benefits my research; Mandi (from ISR), Santanu, Jill, Michael, Aaron, Dan, Carolina, and other fellow JPSM students in the PHD program, they have all contributed to my dissertation writing by providing helpful discussions, encouragement, and invaluable friendship.

The last but not the least, I want to thank my parents and my aunts, who, even remotely living in China, constantly make me feel their dedicated affection as if they were just at my side, and my husband, Tianbing Qi, for the fullness of his love and the entirety of his support towards myself and our family.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction and Literature Review

1.1 The Need for Small Area Estimation

Sample surveys are usually designed to produce estimates for the target survey population and for major population subgroups. Standard survey estimates for major subgroups are termed "design-based" or "direct" estimates because they are based only on the survey data and the selection probabilities for the sample in the subgroup of interest. Statistical inferences based on direct estimates under the usual design-based mode of inference do not depend on the validity of a statistical model, unlike the situation in most other areas of statistics. However, the design-based mode of inference becomes problematic when the sample sizes in the subgroups of interest are small (or even zero). In this situation, model-dependent methods are increasingly being used to produce what are termed as "small area" or "indirect" estimates.

The term "small area" usually refers to a small geographic area such as a state, county, municipality, school district, metropolitan area, or a small domain such as a specific age-sex-race group within a large geographic area. Small areas can be design domains, which are included in the sampling design stage, for example, as strata or primary sampling units (PSUs), or analytic domains which are identified only during the analysis phase of the study.

During the past three decades, the demand for survey estimates for small areas has increased dramatically in many different areas of application, including income and poverty, education, health, substance use, and agriculture. The reason for the increased demand for small area estimates is to be found in the recent trend in federal

policy to target social and economic programs at a more local level. The survey data that were originally designed to provide statistically reliable, design-based estimates of characteristics for a high level of aggregation (e.g., for the nation as a whole, for a large geographic domain such as region), are now also being used to generate model-dependent estimates at a lower level (e.g., states, counties).

Estimates for small areas may be used for allocation of federal funds in government programs, regional planning, and program evaluation. For example, in order to provide updated and precise estimates of income and poverty statistics for the administration of federal programs and the allocation of federal funds to local jurisdictions, the U.S. Census Bureau, with support from other Federal agencies, created the Small Area Income and Poverty Estimates (SAIPE) program beginning in the early 1990's. The program produces timely estimates for several characteristics of income and poverty than the decennial Census in small areas including states, counties, and school districts. Citro and Kalton (2000) reviewed a variety of uses of these estimates. The following summarizes some of them. The estimates produced by the SAIPE program are used to allocate more than $130 billion of U.S. federal funds each year to states and localities. States also use SAIPE estimates to allocate their own and federal funds to substate areas. The Improving America's Schools Act of 1994 called for the use of the SAIPE estimates of poor school-aged children (aged 5-17) for counties and school districts to allocate more than $7 billion (now over $12 billion) of federal funds annually for programs providing extra help to educationally disadvantaged children under Title I of the Elementary and Secondary Education Act.

Estimates of health related statistics in small areas are needed for local health planning and treatment. The U.S. National Health Planning and Resources Development Act of 1974 strongly emphasize local health planning and require local Health System Agencies to collect and analyze data related to the health status of their residents and to the health delivery systems in their health service areas (Nandram, 1999). The U.S. Substance Abuse and Mental Health Services Administration (SAMHSA) produces small area estimates for more than 20 outcomes related to substance use, treatment, and mental health in areas such as states, groups of counties, and census tracts based on data from the National Household Survey on Drug Use and Health (NSDUH) in order to give policy officials a better perspective on the variability in prevalence within and across states. States use these estimates for treatment planning purposes. For more information, see the SAMHSA website (*http://www.oas.samhsa.gov/2k5State/AppA.htm*) and the Research Triangle Institute small area estimation website (*http://www.rti.org/page.cfm/Small_Area_Estimation*).

The Adult Education Amendments of 1988 requires the U.S. Department of Education to submit a report to Congress on the definition of literacy and then report on the nature and extent of literacy among adults in the nation (*http://nces.ed.gov/NAAL/naalhistory.asp*). To satisfy this requirement, the National Center for Education Statistics (NCES) conducted the 1992 National Adult Literacy Survey (NALS) and the 2003 National Assessment of Adult Literacy (NAAL) to assess the English language literacy skills of adults in the U.S. based on an assessment containing a series of literacy tasks completed by sampled adults. These surveys produced direct estimates of English language literacy for the nation and

major subdomains of interest. However, policymakers, researchers, and business leaders often need adult literacy estimates, particularly at the lowest literacy level, for all states and for smaller jurisdictions within the states. This need has led to the production of small area estimates for all states and counties within the U.S. using the survey data from NAAL and NALS (Mohadjer et al., 2007; 2008).

The need for small area estimates is also growing in other countries. For example, policy makers in the U.K. need information about local areas for economic planning, resource allocation and policy making. Recognizing this need, the U.K. Office of National Statistics (ONS) has been carrying out research into the most appropriate ways of constructing small area estimates. A typical example is the Small Area Estimation Project (SAEP) established by the Statistical Methodology Division of the ONS in April, 1998 (SAEP Report, 2003). The Australian Bureau of Statistics (ABS) acknowledges the demand for small area data in various areas in Australia to support planning, decision making and service delivery at local area level. To increase knowledge and understanding of small area estimation techniques, to ensure greater consistency in their application, and to provide a guide for choosing the best method to apply for a particular situation, ABS published a small area estimation practice manual in 2005 (Australian Bureau of Statistics, 2005).

The growing demand for small area estimates applies in many different areas of application and in many countries. Rao (2003) and Jiang and Lahiri (2006a) provide many more examples.

1.2 Direct Estimation for Small Areas

Survey data are extensively used to produce reliable direct estimates of totals or means not only for the population surveyed but also for large areas or domains. A direct estimate for a characteristic of interest in a domain is usually based on the sample units in the domain. Traditional theories on direct domain estimation under the design-based framework are covered in sampling theory books such as Cochran (1977), Sarndal, Swensson and Wretman (1992), and Lohr (1999). Those theories are developed for large domains. In this dissertation, we focus on estimates for small domains, i.e., small areas.

Suppose there are m small areas of interest. Let U_i and s_i denote the index set of the units in a finite population and in a sample, respectively, that are in area i, $i = 1,...,m$. Let N_i and n_i be the number of population units and sample units respectively in area i. Let y_{ik} denote the response for a certain characteristic of interest for the k th unit in the i th small area $(i = 1,...,m;\ k = 1,...,N_i)$. Suppose we want to estimate the population mean $\bar{Y}_i = \sum_{k=1}^{N_i} y_{ik} / N_i$ for the i th small area, as well as the associated variance of the estimator, using the sample drawn from the finite population under a complex sample design. Let w_{ik} denote the sampling weight for sampled unit k in small area i $(i = 1,...,m;\ k = 1,...,n_i)$, which is defined as the inverse of the first-order inclusion probability under the sample design used.

If the sample size in a small area is positive and the sampling weights are the same within the small area, that is, when we have an equal probability of selection

(EPSEM) design within the small area, then the sample mean $\bar{y}_i = \sum_{k=1}^{n_i} y_{ik} / n_i$ is an unbiased estimator of the population mean $\bar{Y}_i$ under the randomization-based inference. When the sampling weights vary within the small area, that is, when we have unequal sampling selection probabilities within the small area, an estimator popularly used among survey practitioners is given by:

$$\bar{y}_{iw} = \frac{\sum_{k=1}^{n_i} w_{ik} y_{ik}}{\sum_{k=1}^{n_i} w_{ik}}, \quad i = 1, ..., m. \tag{1.1}$$

This estimator was proposed by Brewer (1963) and Hajek (1971). The traditional design-based domain estimation techniques developed in sample surveys (e.g., Cochran, 1977; Sarndal, Swensson and Wretman, 1992; Lohr, 1999) and resampling methods (Wolter, 1985) may also be used to estimate the associated sampling variances for the direct small area means.

However, these design-based estimates (or direct estimates) are very imprecise when the sample sizes in the small areas are small, or are even unavailable when the sample size is zero. Therefore, alternative approaches have to be used in order to produce reliable estimates for small areas of interest. The demand for precise small area estimates has led to the development of model-dependent techniques of small area estimation (SAE).

1.3 Model-based Estimation for Small Areas

In the absence of adequate sample sizes, any improved estimation procedure calls for statistical models that can combine information from related sources. Small

area estimation techniques combine information from a variety of relevant sources to form indirect estimators that generally increase the precision of the small area estimates. These indirect estimators are based on various implicit or explicit models that provide a link to related small areas through supplementary data (e.g., recent census and/or administrative records).

A variety of indirect estimators have been proposed in the literature. One of the first was synthetic estimation (Gonzales, 1973; Gonzales and Hoza, 1978). However, this methodology produces model-unbiased estimators under a very restrictive model, which is usually unrealistically simple. Composite estimators were developed to balance the potential bias of the synthetic estimator under model failure against the instability of a direct estimator by taking a weighted average of the two estimators. One main challenge in composite estimation is how to determine the weight. Later on, model-based composite estimators based on realistic or explicit small area models that account for local variations were developed (Rao, 2003). An explicit model is useful since it gives users an idea of the data generation process and how different information sources are combined. Among the range of explicit small area models, mixed models that include both fixed effects and random area-specific effects have been widely used in small area estimation in recent years (Jiang and Lahiri, 2006a). We briefly review them in the next section.

1.4 Mixed Models in Small Area Estimation

The popularity of mixed models in SAE is owing to their flexibility in combining information from different sources and taking account of different sources

of error. A mixed model typically incorporates area-specific random effects that reflect additional between-area variations in the data that are not explained by the fixed effects part of the model. Two primary types of mixed model have been employed in the small area estimation literature: area level models and unit level models.

1.4.1 Area Level Model

A general area (or aggregate) level model consists of two models. One is the sampling model that accounts for the sampling error of the direct survey estimates. The other is the linking model that relates the population value to a set of known area-specific auxiliary variables. Since the design-based survey estimates are modeled directly, area level models usually produce design-consistent estimators. However, area level models require precise estimates of the sampling variances of the design-based survey estimates, which is a challenging problem due to the small sample sizes in the small areas.

Basic Area Level Model

A typical example of a basic area level model is the Fay-Herriot model (Fay and Herriot, 1979). Assume that $\theta_i = h(\overline{Y}_i)$ is related to area-specific auxiliary data $\mathbf{z}_i = \left(z_{i1},...,z_{ip}\right)'$ through a linear model for some specified function $h(\bullet)$. Let $\hat{\theta}_i$ denote the direct estimate of θ_i. In order to estimate the per-capita income (PCI) in

1969 for small places (i.e., population less than 1000) in the United States, Fay and Herriot (1979) proposed the following two-level model, often referred to as the *Fay-Herriot model*:

$$\text{Level 1 (sampling model): } \hat{\theta}_i \mid \theta_i \overset{ind}{\sim} N(\theta_i, \psi_i), \quad i = 1, ..., m, \tag{1.2}$$

$$\text{Level 2 (linking model): } \theta_i \mid \boldsymbol{\beta}, \sigma_v^2 \overset{ind}{\sim} N(\mathbf{z}_i'\boldsymbol{\beta}, \sigma_v^2), \; i = 1, ..., m, \tag{1.3}$$

where level 1 is used to account for the sampling variability of the regular survey estimates $\hat{\theta}_i$ of the true small area means θ_i; level 2 links θ_i to a vector of p known auxiliary variables $\mathbf{z}_i$. The two-level model (1.2)-(1.3) is also referred to as a *matched* model because (1.2)-(1.3) can be combined into a single linear mixed model with the following form:

$$\hat{\theta}_i = \mathbf{z}_i'\boldsymbol{\beta} + v_i + e_i, \quad i = 1, ..., m, \tag{1.4}$$

where v_i is the random area effect and e_i is the sampling error. Further, $v_i \overset{iid}{\sim} N(0, \sigma_v^2)$ and $e_i \overset{ind}{\sim} N(0, \psi_i)$ are commonly assumed. Under model (1.4), the true small area mean can be written as $\theta_i = \mathbf{z}_i'\boldsymbol{\beta} + v_i$. The parameters $\boldsymbol{\beta}$ and σ_v^2 are generally unknown and are estimated from the available data. The sampling variances ψ_i are customarily assumed known, whereas in practice, they have to be estimated. A commonly used approach is to estimate ψ_i from the unit-level data using the traditional domain estimation techniques first, and then smooth the estimated variances $\hat{\psi}_i$ to get more stable estimates of ψ_i. Fay and Herriot (1979) used logarithmic transformation $\hat{\theta}_i = \log(\bar{y}_i)$ in order to stabilize the sampling variance.

More recently, a generalized variance function (GVF) technique (Wolter, 1985; Valliant, 1987) has been commonly employed to smooth the sampling variances in the small area estimation problem (e.g., see Otto and Bell, 1995; Mohadjer et al., 2007). GVFs estimate variances (or relative variances) through suitable models which describe the relationship between the variance (or relative variance) of a survey estimator and its expectation.

The Fay-Herriot type of model with different choices of transformation function $h(\bullet)$ has been extensively used in small area estimation and related problems by practitioners and researchers. A recent example is the SAIPE state level model. SAIPE has applied the Fay-Herriot type of model without transformation (i.e., $\hat{\theta}_i = \bar{y}_i$) to produce mean household income estimates and poverty rates by age group for all U.S. states since 1993 (Citro and Kalton, 2000; Maples and Bell, 2005).

Prior to Fay and Herriot (1979), Efron and Morris (1975) applied the arc-sine transformation [$\hat{\theta}_i = \sqrt{n_i}\arcsin(2p_i - 1)$] to the sample proportions p_i in order to stabilize the sampling variance in their well-known baseball data example, where they used the two-level Fay-Herriot model (1.2)-(1.3) without any covariates to predict the batting average for all the players for the remainder of the 1970 season based on their batting averages for the first 45 at bats, p_i (sampling proportion). Carter and Rolph (1974) applied a similar transformation function [$\hat{\theta}_i = \arcsin\left(\sqrt{p_i}\right)$] in their false alarm probability estimation example.

Following Fay and Herriot (1979), SAIPE has also applied the logarithmic transformation for their county level model in estimating poverty rates or counts of

school-age children for all U.S. counties (Citro and Kalton, 2000). To overcome the problem that the logarithmic transformation cannot be applied to the areas with zero direct estimates, Fisher and Asher (1999, 2000) developed a hierarchical Bayes model based on a scaled binomial kernel as an alternative to the SAIPE county models.

The justification for the transformation method is based on the central limit theorem, which relies on the sample size being large. That is, a well chosen transformation of a direct estimate will be more nearly normally distributed than the direct estimate itself. If the area level sample size is small, this approach is less effective.

Extensions of the Basic Area Level Model

Various extensions of the basic area level model have been developed in the literature. One type of extension is to extend the univariate model (1.4) to a multivariate model in order to take advantage of the correlations between different characteristics of interest. For example, Fay (1987) and Datta, Fay and Ghosh (1991) considered the following multivariate model as an extension to model (1.4):

$$\hat{\boldsymbol{\theta}}_i = \mathbf{Z}_i \boldsymbol{\beta} + \mathbf{v}_i + \mathbf{e}_i, \quad i = 1, ..., m, \tag{1.5}$$

where $\hat{\boldsymbol{\theta}}_i = (\hat{\theta}_{i1}, ..., \hat{\theta}_{ir})'$ is an $r \times 1$ vector of the direct estimates of the characteristics of interest, $\mathbf{Z}_i$ is an $r \times rp$ matrix with j th row given by $\left(\mathbf{0}', ..., \mathbf{0}', \mathbf{z}_{ij}', \mathbf{0}', ..., \mathbf{0}'\right)$ with $\mathbf{0}$ being the $p \times 1$ null vector, $\boldsymbol{\beta}$ is the $rp-$vector of regression coefficients, $\mathbf{v}_i$ are the area-specific random effects which are assumed to be independent multivariate

normal with mean $\mathbf{0}$ (the $r \times 1$ null vector) and variance $\boldsymbol{\Sigma}_v$, i.e., $\mathbf{v}_i \overset{ind}{\sim} N_r\left(\mathbf{0},\ \boldsymbol{\Sigma}_v\right)$, and $\mathbf{e}_i = (e_{i1},...,e_{ir})'$ are the sampling errors which are assumed to be independent multivariate normal with mean $\mathbf{0}$ and known covariance matrix $\boldsymbol{\Psi}_i$, i.e., $\mathbf{v}_i \overset{ind}{\sim} N_r\left(\mathbf{0},\ \boldsymbol{\Psi}_i\right)$. The authors demonstrated that the multivariate model (1.5) can lead to more efficient estimators of the small area means than the univariate model (1.4). Note that model (1.5) is general enough to allow the vector of covariates to be different for every characteristic, although, in practice, datasets may not be rich enough to support this.

A second type of extension is to extend the basic area level model to a model that can handle cross-sectional and time series data (e.g., the labor force survey data). For instance, a cross-sectional and time series model with the following form has been considered in the literature:

$$\hat{\theta}_{it} = \mathbf{z}_{it}'\boldsymbol{\beta} + v_i + u_{it} + e_{it}, \quad i = 1,...,m;\ t = 1,...,T, \tag{1.6}$$

where $\hat{\theta}_{it}$ is the direct survey estimate of the characteristic of interest for small area i at time t, $v_i \overset{iid}{\sim} N\left(0,\ \sigma_v^2\right)$ are the area-specific random effects, $\mathbf{z}_{it}$ is a vector of area-specific covariates, some of which may change with time t, e_{it} are sampling errors normally distributed with zero means and a known block diagonal covariance matrix $\boldsymbol{\Psi}$ with blocks $\boldsymbol{\Psi}_i$. Rao and Yu (1994) proposed a model with form (1.6) assuming that the u_{it} follow a common first-order autoregressive model (AR1) for each i. Datta, Lahiri and Maiti (2002) and You (2008) considered models similar to (1.6)

assuming that the u_{it} follow a random walk model. Datta et al. (1999) considered a model similar to (1.6), adding extra terms to the linking model to reflect seasonal variation in their application of estimating unemployment rates for all U.S. states. You, Rao and Gambino (2003) employed a simpler cross-sectional and time-series model than the one developed by Datta et al. (1999) to produce small area estimates of unemployment rates for the Canadian Labor Force Survey. For more versions and applications of the cross-sectional and time series model, we refer to Section 5.4.3 of Rao (2003).

A third type of extension is to extend the basic area level model to *unmatched* sampling and linking models. When θ_i is not a linear function of $\overline{Y}_i$, the sampling error assumption $E\left(e_i \mid \theta_i\right) = 0$ in model (1.4) may not be valid for areas with small sample sizes (Rao, 2003, Sec. 5.2 and 10.4). To overcome this problem, You and Rao (2002a) proposed an unmatched sampling and linking model with the following form:

$$\text{Level 1 (sampling model): } \overline{y}_i \mid \theta_i \overset{ind}{\sim} N(\theta_i,\ \psi_i),\ i = 1,...,m, \tag{1.7}$$

$$\text{Level 2 (linking model): } g(\theta_i) \mid \boldsymbol{\beta}, \sigma_v^2 \overset{ind}{\sim} N(\mathbf{x}_i'\boldsymbol{\beta}, \sigma_v^2),\ i = 1,...,m, \tag{1.8}$$

where $\overline{y}_i$ is a design-unbiased survey estimate of the small area mean θ_i, and $g\left(\bullet\right)$ is a specific linking function. The *logarithm* and *logit* functions are commonly used as linking functions. The two-level model defined by (1.7) and (1.8) is called an unmatched model in the sense that the sampling and linking models cannot be combined into a single linear mixed model. You and Rao (2002a) applied their proposed unmatched model (1.7)~(1.8) with $g\left(\theta_i\right) = \log\left(\theta_i\right)$ to the estimation of Canadian Census under-coverage.

Following You and Rao (2002a), Mohadjer et al. (2007) developed an unmatched area-level model which incorporates both state and county random effects to produce estimates of the percentages of adults at the lowest level of English language literacy for all states and counties in U.S. using the 2003 NAAL data. The *logit* function was used as the linking function in order to guarantee the outcome of the small area estimates falling into the right range of (0, 1).

A fourth type of extension is to replace the normality assumptions typically assumed for the basic area level model by some more appropriate alternatives. The typical Fay-Herriot type of model assumes normality for the error components in the models, namely the area-level random effects and/or the sampling errors of the direct survey estimates. However, real data often show significant departures from normal distributions. Heavy-tailed distributions and asymmetric distributions are frequently encountered in empirical studies (Sec. 1.2 of Hampel et al., 1986; Azzalini, 1985; 1986). For cases where the assumption of normality is not tenable, more flexible models can be adopted to accommodate non-normal features related to skewness, kurtosis, and heavy tails. However, the literature in small area estimation on this aspect is not rich.

Some researchers have extended the Fay-Herriot models by assuming non-normal distributions for the area-level random effects as a means of dealing with outliers. For example, Datta and Lahiri (1995) developed a model assuming that the random effect follows a scale mixture of normal distributions, where the $t-$distribution is a special case. Huang and Bell (2006) proposed an extension of the SAIPE Fay-Herriot state level model by assuming either the random effects or the

sampling errors (but not both) follow a $t-$distribution. Xie et al. (2007) used an extension of the Fay-Herriot model by assuming the random area effects follow a $t-$distribution with unknown degrees of freedom in order to produce estimates of the proportion of overweight individuals in small areas using the 2003 public-use Behavioral Risk Factor Surveillance System (BRFSS) data. Recently, Fabrizi and Trivisano (2007) proposed two extensions of the Fay-Herriot model by assuming the random area effects follow either an exponential power distribution or a skewed exponential power distribution.

For more extensions of the basic area level mixed model and their applications, we refer to Rao (2003).

1.4.2 Unit Level Model

A unit (or respondent) level mixed model can be used when unit-specific response variables are available in each small area. This class of models can incorporate auxiliary information at both the unit and area level (Moura and Holt, 1999). The area-specific random effect terms in unit level models can capture the correlation possibly present among the sample units within a small area. The main advantage of unit level models is that they can incorporate all sources of uncertainty; in particular, they can capture the uncertainty due to the estimation of the sampling variances.

Design-consistent model-based estimators are appealing to survey practitioners because such estimators provide protection against model failures as the small area sample sizes increase (Rao, 2003, p. 148). As we mentioned in Section

1.4.1, estimators produced using area level models usually satisfy the design-consistency property because area level models employ the design-unbiased direct estimates which account for the survey design in the sampling model. Unit level models do not use design-based estimators directly. In order to produce design-consistent estimators using a unit level model, if the detailed design information (e.g., stratification and clustering) is available at the individual level, one can build a unit level model which incorporates all the design information, although the modeling may become challenging if a very complex design is used. If the design information is not available at the individual level, one can incorporate the survey weights following the approaches outlined in Kott (1989), Prasad and Rao (1999), You and Rao (2002b). The fundamental idea is to obtain a survey-weighted aggregated area level model from the unit level model by taking a weighted average with weights being normalized survey weights.

We now briefly review two primary types of unit level mixed model employed in the SAE literature: the unit level linear mixed model and the unit level generalized linear mixed model.

Unit Level Linear Mixed Model

A simple example of a unit level mixed model is the nested error regression model originally employed by Battese, Harter and Fuller (1988) to estimate areas under corn and soybean for each of the 12 counties of North Central Iowa using survey and satellite data. They used the following linear mixed model (BHF model):

$$y_{ik} = \mathbf{x}_{ik}'\boldsymbol{\beta} + v_i + e_{ik}, \quad k = 1,...,n_i; \ i = 1,...,m, \tag{1.9}$$

where y_{ik} is the number of hectares of corn (or soybeans) in the k th segment of the i th county, and the random error terms v_i and e_{ik} are assumed i.i.d $N(0,\ \sigma_v^2)$ and $N(0,\ \psi)$. The random term v_i represents the effect of area characteristics that are not accounted for by the auxiliary variables $\mathbf{x}_{ik}$. Under model (1.9), the true small area mean θ_i can be written as $\theta_i = \overline{\mathbf{x}}'_{i(p)}\boldsymbol{\beta} + v_i$, where $\overline{\mathbf{x}}_{i(p)} = \Sigma_{k=1}^{N_i}\mathbf{x}_{ik} / N_i$ and N_i is the total number of segments in the i th county.

Model (1.9) can produce design-consistent estimators only for data collected using a simple survey design. More complex models are needed to handle data collected from complex survey designs. For data collected from a stratified two-stage sampling, where the strata were the small areas, Stukel and Rao (1999) proposed the following two-fold nested error regression model:

$$y_{ijk} = \mathbf{x}'_{ijk}\boldsymbol{\beta} + v_i + u_{ij} + e_{ijk},\ k = 1,...,N_{ij};\ j = 1,...,M_i;\ i = 1,...,m, \qquad (1.10)$$

where y_{ijk} and $\mathbf{x}_{ijk}$ are the response values of the characteristic of interest and the associated auxiliary variables for individual k in primary sampling unit (PSU) j in small area i respectively; $v_i \overset{iid}{\sim} N(0,\sigma_v^2)$ are the random area-specific effects; $u_{ij} \overset{iid}{\sim} N(0,\sigma_u^2)$ are the random within area PSU effects, and $e_{ijk} \overset{iid}{\sim} N(0,\sigma_e^2)$ are the sampling errors. Ghosh and Lahiri (1988) studied model (1.10) for the case of no auxiliary information, i.e., $\mathbf{x}'_{ijk}\boldsymbol{\beta} = \beta$, and did not specify a specific distribution for v_i and u_{ij} as a robustness feature of their method. Datta and Ghosh (1991) used model (1.10) for the special case of cluster-specific covariates, that is, $\mathbf{x}_{ijk} = \mathbf{x}_{ij}$.

Logistic Regression Model with Mixed Effects

Linear mixed models like the BHF and the two-fold nested error regression models are applicable for continuous observations. Recent research in SAE focuses on the situations where the dependent variables are categorical or discrete and where the small area parameters of interest are proportions or counts. In such cases, a unit level linear mixed model is no longer applicable. Generalized linear mixed models (GLMM) are thus developed to fulfill the needs (McCulloch, 2003; Jiang and Lahiri, 2006a). Among those, logistic regression models with mixed effects are commonly used models in estimating small-area proportions.

To estimate the census undercount for local areas, Dempster and Tomberlin (1980) proposed an empirical Bayes method based on a logistic regression model containing both fixed and random effects. This proposal was further developed by MacGibbon and Tomberlin (1989). In order to estimate the true area proportions $P_i = \Sigma_{k=1}^{N_i} y_{ik} / N_i$, MacGibbon and Tomberlin (1989) proposed the following model:

$$\begin{aligned} & y_{ik} \mid p_{ik} \overset{ind}{\sim} Bernoulli(p_{ik}), \\ & \text{logit}(p_{ik}) = \log\left[p_{ik} / (1 - p_{ik}) \right] = \mathbf{x}'_{ik}\boldsymbol{\beta} + v_i; \; v_i \overset{iid}{\sim} N(0, \sigma_v^2), \end{aligned} \tag{1.11}$$

where p_{ik} denotes the probability of a response for the k th unit in the i th area, and y_{ik} and $\mathbf{x}_{ik}$, $k = 1,...,n_i;\; i = 1,...,m$, are unit-specific binary (0 or 1) responses of the characteristic of interest and covariates respectively. The model-based estimator of P_i was obtained using $\hat{p}_i = \Sigma_{k=1}^{N_i} \hat{p}_{ik} / N_i$, where $\hat{p}_{ik}$ is obtained from (1.11) by estimating $\boldsymbol{\beta}$ and the realization of v_i through empirical Bayes or hierarchical Bayes

methods which will be discussed in Section 1.5. Applications of similar models can also be found in Wong and Mason (1985) and Tomberlin (1988). Farrell, MacGibbon and Tomberlin (1997a and b) further developed the model to produce estimates of small area proportions in multistage designs.

Malec et al. (1997) considered a different logistic regression model with random regression coefficients using data from the National Health Interview Survey (NHIS), a multistage, personal interview sample survey that is conducted annually by the National Center for Health Statistics. Suppose each individual in the population is assigned to one of J mutually exclusive and exhaustive classes based on the individual's socioeconomic/demographic status. The binary response y_{ijk} for individual k $(k = 1,..., N_{ij})$ in class j in cluster i is assumed independent *Bernoulli* with common probability p_{ij}. In the absence of detailed design information, to make inferences about a finite population proportion for a specified small area and subgroup $P = \sum_{i\in I}\sum_{j\in J}\sum_{k=1}^{N_{ij}} y_{ijl} \Big/ \sum_{i\in I}\sum_{j\in J} N_{ij}$, where I is the collection of clusters that define the small area and J is the collection of classes that defines the subpopulation, the following models are assumed:

$$
\begin{aligned}
& y_{ijk} \mid p_{ij} \overset{ind}{\sim} Bernoulli(p_{ij}); \\
& \text{logit}(p_{ij}) = \log\left[p_{ij} \Big/ \left(1 - p_{ij}\right)\right] = \mathbf{x}'_j \boldsymbol{\beta}_i; \\
& \boldsymbol{\beta}_i = \mathbf{Z}_i \alpha + \mathbf{v}_i; \quad \mathbf{v}_i \overset{iid}{\sim} N(0, \boldsymbol{\Sigma}_v);
\end{aligned}
\tag{1.12}
$$

where $\mathbf{x}_j$ is the class-specific covariate vector and $\mathbf{Z}_i$ is a $p \times q$ area level covariate matrix. Note that the first level of (1.12) can be transformed to a binomial

model $y_{ij} \mid p_{ij} \overset{ind}{\sim} Binomial\left(p_{ij},\ n_{ij}\right)$, and thus model (1.12) can also be considered as an area level model. We also note that model (1.12) does not take into account the detailed design information of NHIS because design information was not avaliable. Thus, the design-consistency of the HB estimators produced using this model is questionable.

Other applications of logistic regression models with mixed effects in estimating small area proportions can be found in Stroud (1991), Malec, Sedransk and Tompkins (1993), Malec, Davis and Cao (1999), Jiang and Lahiri (2001), among others. GLMM also includes models for mortality and disease rates, exponential family models, semi-parametric models, etc. For details of these applications and for other references, we refer to Pfeffermann (2002) and Rao (2003).

1.5 Inference Using Mixed Models

Based on the mixed models reviewed in Section 1.4, small area estimates can be expressed as a linear or nonlinear combination of the fixed and random effects. Two primary approaches – the empirical best prediction (EBP) approach and hierarchical Bayesian (HB) approach, have been used for inference about the small area parameters (e.g., means, totals, proportions, etc.) using mixed models. Both approaches are used to approximate the conditional probability distributions that arise from Bayes' theorem for the small area quantities. The EBP approach is usually referred to as a classical (or frequentist) approach because it uses classical methods to estimate the unknown hyperparameters of the mixed model under consideration (e.g.,

$\boldsymbol{\beta}$ and σ_v^2 in the Fay-Herriot model). The HB approach is referred to as the Bayesian approach because it assumes prior distributions on the unknown hyperparameters. Rao (2003) reviews both approaches. Jiang and Lahiri (2006a) provided an extensive review of the EBP approach. We briefly review the two approaches in the following two subsections.

1.5.1 Empirical Best Prediction Approach

Consider the area level model (1.4). Under that model, the small area mean θ_i can be written as $\theta_i = \mathbf{x}_i'\boldsymbol{\beta} + v_i$. Assuming that both $\boldsymbol{\beta}$ and σ_v^2 are known, we can obtain the best predictor (BP) of θ_i in the form:

$$\theta_i^{BP} = \mathbf{x}_i'\boldsymbol{\beta} + (1 - B_i)(y_i - \mathbf{x}_i'\boldsymbol{\beta}) \text{, where } B_i = \psi_i \Big/ \left(\sigma_v^2 + \psi_i\right),\ i = 1,...,m.$$

Next, assume σ_v^2 is known and $\boldsymbol{\beta}$ is unknown. Let $\mathbf{X} = (\mathbf{x}_1,...,\mathbf{x}_m)'$, $\mathbf{Y} = (y_1,...,y_m)'$, $\mathbf{V} = Diag(\sigma_v^2 + \psi_1,...,\sigma_v^2 + \psi_m)$. We can get the maximum likelihood estimator (MLE) of $\boldsymbol{\beta}$ in the form: $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}$. Replacing $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}}$ in the BP, we can get the best linear unbiased predictor (BLUP) of θ_i with the form: $\theta_i^{BLUP} = \mathbf{x}_i'\hat{\boldsymbol{\beta}} + (1 - B_i)(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}})$. In practice, the variance components for the random effects (namely, parameter σ_v^2 in the model (1.4)) are rarely known. A common procedure under the EBP approach is to replace them in the BLUP by the standard variance component estimators obtained through the methods of moments (Fay and Herriot, 1979; Prasad and Rao, 1990), maximum likelihood (ML), or

restricted maximum likelihood (REML). The resulting estimator of θ_i is called empirical BLUP (EBLUP). Common limitation of these methods is the possibility of producing zero or negative estimates of the variance components. The Adjusted Density Maximization (ADM) methods proposed by Morris (1988) and Li (2007) are very promising because they avoid zero or negative estimates of the variance components and have good asymptotic properties. For an extensive review on different variance component estimation methods, we refer to Jiang and Lahiri (2006a).

Estimation of the variance components is relatively easy, whereas, assessment of the uncertainty due to the estimation is quite challenging. Extensive references can be found in the SAE literature for assessing this uncertainty. For instances, Prasad and Rao (1990) used a mean squared error (MSE) criterion to measure the uncertainty of EBLUP under a general linear longitudinal mixed model, and proposed a second-order approximation to the MSE using the Taylor series method under normality assumption for the Fay-Herriot model. Lahiri and Rao (1995) demonstrated the robustness of this approximation against non-normality. Kleffe and Rao (1992) provided a second-order approximation to the MSE of EBLUP using a random error variance linear model. Their work was further extended by Butar and Lahiri (2002) to a more general model. More recently, Jackknife methods (e.g., see Jiang, Lahiri and Wan, 2002; Chen, 2001; Lohr and Rao, 2007) and parametric bootstrap methods (e.g., see Butar and Lahiri, 2003; Lahiri, 2003; Pfeffermann and Tiller, 2005; Hall and Maiti, 2006) have been proposed to estimate the MSE of EBLUP. Chatterjee, Lahiri and Li (2008) used a parametric bootstrap approximation to study the distribution of

EBLUP and related prediction intervals, a quite challenging problem under the EBP approach. For extensive reviews on these resampling methods, we refer to the review paper by Gershunskaya, Jiang and Lahiri (2008).

1.5.2 Hierarchical Bayesian Approach

In the HB approach, a subjective prior distribution on the hyperparameters is specified and the posterior distribution of the parameter of interest is obtained. The HB approach is straightforward compared to EBP in the sense that, the posterior distributions, once computed, can be used for all inferential purposes. A second advantage of the HB approach is its ability to incorporate complex models which EBP approach cannot handle easily, such as unmatched sampling and linking models (e.g., see You and Rao, 2002a), and models assuming the random effects follow a class of distributions in stead of relying on the normal distribution (e.g., see Datta and Lahiri, 1995; Fabrizi and Trivisano, 2007). Another advantage of the HB approach is its flexibility to take account of the uncertainty of the direct sampling variances by assuming prior distributions. As we mentioned earlier, the direct sampling variances are assumed known in area-level mixed models even though they are often estimated and smoothed in practice using techniques such as GVF. This extra uncertainty was not assessed by any of the EBP approach until the work by Arora and Lahiri (1997). Hinrich (2003) expanded the Arora and Lahiri model and proposed some new models to account for the uncertainty of the sampling error using area-level mixed models.

HB inferences can be implemented using the Markov Chain Monte Carlo (MCMC) technique. MCMC methods are a class of algorithms for sampling from a

probability distribution by constructing and simulating a Markov chain that has the desired distribution as its equilibrium distribution. Robert and Casella (1999) and Rao (2003, Sec. 10.2) describes MCMC methods in detail. We briefly review them here. Let $\boldsymbol{\eta} = (\boldsymbol{\theta}, \boldsymbol{\lambda})'$ be the vector of small area parameters $\boldsymbol{\theta}$ and model parameters $\boldsymbol{\lambda}$. In general, it is not feasible to draw independent samples from the joint posterior distribution $f(\boldsymbol{\eta} \mid \mathbf{y}_s)$, because the denominator of the posterior, $f_1(\mathbf{y}_s)$, is usually intractable. MCMC avoids this difficulty by constructing a Markov chain $\left\{\boldsymbol{\eta}^{(k)},\ k = 0,\ 1,\ 2,\ ...\right\}$ with a starting point $\boldsymbol{\eta}^{(0)}$ such that the distribution of $\boldsymbol{\eta}^{(k)}$ converges to a unique stationary distribution, $\pi(\eta)$, which is equivalent to the posterior distribution $f(\boldsymbol{\eta} \mid \mathbf{y}_s)$. Therefore, after a sufficiently large "burn in", d, we can treat $\boldsymbol{\eta}^{(d+1)},...,\boldsymbol{\eta}^{(d+T)}$ as T dependent samples from the target distribution $f(\boldsymbol{\eta} \mid \mathbf{y}_s)$, regardless of the starting point. The average of the sequence $\left\{\boldsymbol{\eta}^{(d+1)},...,\boldsymbol{\eta}^{(d+T)}\right\}$ can be used to approximate the posterior mean $E(\boldsymbol{\eta} \mid y_s)$. This property follows from the ergodic theorem of stochastic process, which can be viewed as the law of large numbers for a dependent sequence.

HB methods are now widely used, largely due to advances in computing power and user-friendly software. Among the broad range of MCMC simulation methods, one algorithm, Gibbs sampling, has been increasingly used in applied Bayesian analyses (see Gelfand and Smith, 1990; Gilks et al., 1996; Robert and Casella, 1999). The appeal of Gibbs sampling is that it can be used to estimate posterior distributions by drawing sample values randomly from the full conditional distributions for each of the individual parameters (i.e., the conditional distribution of

a parameter given the other parameters and the observed data). On many occasions, the full conditional distributions do not have closed form; in such cases, some rejection sampling algorithm, such as the Metropolis-Hastings (M-H) algorithm within the Gibbs sampler, can be used (Chib & Greenberg, 1995). The necessary computation routines are now freely available in the software package WinBUGS (*http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/dicpage.shtml*), which makes the implementation of Bayesian methods straightforward.

In addition to MCMC, HB estimation can also be implemented using alternative approximate methods including Laplace's method (e.g., see Laplace, 1847; Erdelyi, 1956), and Gauss-Hermite Quadrature (e.g., see Davis and Rabinowitz, 1975). Applications of Laplace's method to approximate a complex posterior distribution and its moments have been explored by a number of researchers, including Tierney and Kadane (1986), Kass, Tierney and Kadane (1988), Tierney, Kass and Kadane (1989), Morris (1988, 2006), Kass and Steffey (1989), Wolfinger (1993), and Christiansen and Morris (1997). The approximations offer simple interpretations of the Bayesian methodology. Some researchers have also applied Laplace's method to obtain the maximum likelihood estimator (MLE) of model parameters (e.g., see Raudenbush et al., 2000; Olsen and Schafer, 2001). Gauss-Hermite Quadrature is a standard numerical approach to approximate integrals. It is often used for numerical integration in statistics because of its relation to Gaussian densities (Liu and Pierce, 1994). Raudenbush et al. (2000) considered Gauss-Hermite Quadrature in addition to Laplace's method to obtain their MLEs of model parameters.

1.6 Auxiliary Data and Model Selection for Small Area Estimation

Auxiliary data (or predictor variables) play an important role in small area estimation. The choice of small area models depends on the availability of auxiliary data and the relationship between these data and the variables of interest at the small area level. Auxiliary data are often obtained from various administrative and census records. In essence, we want to "borrow strength" from these auxiliary data to increase the accuracy of the estimates for small areas.

When a large pool of potential auxiliary variables is available, the selection of a smaller set of suitable auxiliary variables is necessary in many small area estimation projects. For instance, only seven auxiliary variables were finally chosen from over 100 potential auxiliary variables in the NAAL small area estimation model (Mohadjer et al., 2008).

Model selection techniques may be applied to select the best set of auxiliary variables given a big pool of potential auxiliary variables. With the classical modeling approach, the Akaike information criterion (AIC) and Bayesian information criterion (BIC, also known as the Schwarz criterion) are commonly used criteria for model selection purposes. For detailed information on these criteria and how to use them in the small area estimation context, we refer to Rao (2003, p. 105-107). With the HB modeling approach, a commonly used model selection criterion is the deviance information criterion (DIC) proposed by Spiegelhalter et al. (2002). Since this dissertation emphasizes HB modeling, we provide details about the DIC in the next few paragraphs.

The DIC is a generalization of the AIC and BIC for a hierarchical model. It is particularly useful in Bayesian model selection problems where the posterior distributions of the models have been obtained by MCMC simulation. Like AIC and BIC, it is an asymptotic approximation as the sample size becomes large. It is valid when the posterior distribution is approximately multivariate normal. DICs are comparable only over models with exactly the same observed data.

Let $\boldsymbol{\theta}$, $\mathbf{y}$, and $p(\mathbf{y}\,|\,\boldsymbol{\theta})$ denote the unknown parameters of the model, the data, and the likelihood respectively. Let $f(\mathbf{y})$ be some fully specified standardizing term that is a function of the data alone. Then define the *deviance* as:

$$D(\boldsymbol{\theta}) = -2\log\big[p(\mathbf{y}\,|\,\boldsymbol{\theta})\big] + 2\log\big[f(\mathbf{y})\big]. \tag{1.13}$$

The second term in the deviance involves $\mathbf{y}$ only and cancels out when comparing deviances for different models; the term can therefore be dropped. The posterior mean of the deviance, $\bar{D} = E\big[D(\boldsymbol{\theta})\,|\,\mathbf{y}\big]$, is a measure of goodness-of-fit of the model; the larger $\bar{D}$, the poorer is the fit. The $\bar{D}$ statistic has been used to compare models in the literature, but this measure does not penalize overly complex models. As the number of parameters in a model increases, $\bar{D}$ decreases.

The measure of the *effective number of parameters* of a Bayesian model is computed as:

$$p_D = \bar{D} - D(\bar{\boldsymbol{\theta}}),$$

where $\bar{\boldsymbol{\theta}} = E(\boldsymbol{\theta}\,|\,\mathbf{y})$ is the posterior mean of the parameters. The measure p_D represents the effect of model fitting. The larger p_D, the easier it is for the model to fit the data.

The DIC is calculated as:

$$DIC = p_D + \bar{D} = 2\bar{D} - D(\bar{\boldsymbol{\theta}}). \tag{1.14}$$

The model with the smallest DIC is judged to be the model that would best predict a replicate dataset of the same structure as the one currently observed. Incorporating p_D in the DIC calculation penalizes models with larger numbers of parameters. The DIC assumes that the posterior mean is a good measure of the stochastic parameter. If this assumption is violated, say because of extreme skewness or even bimodality, then the DIC may not be appropriate. For more details on the DIC, we refer to Spiegelhalter et al. (2002) and Gelman et al. (2004).

Even though DIC is a useful model selection criterion, implementing a HB model with large number of auxiliary variables may be quite challenging. The covariance matrix may be singular, and the convergence speed may be slow because of the large number of model parameters. In such situations, classical stepwise regression may be implemented as a preliminary step for selecting auxiliary variables among a large pool of potential auxiliary variables for the HB models (e.g., see Malec et al., 1997; Mohadjer et al., 2007).

Jiang et al. (2008) recently introduced a new class of strategies, known as fence methods, for mixed model selection. The models include linear and generalized linear mixed models. Unlike AIC, BIC and DIC, fence methods do not try to minimize a criterion function. The optimization procedure of fence methods involves two steps. The first step is to isolate a subgroup of correct models, including the optimal model, by constructing a statistical fence to carefully eliminate incorrect

models. The second step is to select the optimal model among those within the fence according to a criterion which can be made flexible.

1.7 Model-based Prediction Methods under Finite Population Sampling

Under the finite population framework, we need essentially to make inferences about the small finite population means $\overline{Y}_i$ based on inferences about the model parameters θ_i. A prediction approach is needed to produce values for the non-sampled units utilizing the small area model employed. Model-based prediction methods under finite population sampling have a long history. We briefly review them in this section.

The model-based approach in survey sampling theory views the finite population as a realization from a hypothetical super-population (Cochran, 1939). Brewer (1963) and Royall (1970) used a prediction approach to estimate the population mean, partly motivated by a super-population model. Under the prediction approach, the super-population model is used to predict values for the non-sampled units from the knowledge gained through the sample. The books by Bolfarine and Zacks (1999) and Valliant et al. (2000), and the review paper by Graubard and Korn (2002) gave comprehensive reviews on this subject. Ghosh and Meeden (1997) explored related Bayes and empirical Bayes approaches. Rao (2005) examined the interplay between sample survey theory and practice over the past 60 years or so.

A concern about the prediction approach is that the predictions may be unreliable in the case of model misspecification. Therefore, model robustness is important, and is studied by researchers from different perspectives. Valliant (1985; 1986) extended Royall's (1970; 1976) super-population approach to cover certain nonlinear models. Li and Lahiri (2007) proposed a new robust prediction approach in which the super-population model is chosen adaptively from the well-known Box–Cox class of probability distributions.

In the Bayesian approach, a prior distribution is assumed on the parameters of the finite population. Ericson (1969) first formulated the normal theory of Bayesian analysis for finite population sampling. Ghosh and Meeden (1986) introduced empirical Bayes estimation of the finite population mean assuming a normal superpopulation model. Ghosh and Lahiri (1987a; b) relaxed the normality assumption and motivated their estimators of means and variances from stratified samples using a linear empirical Bayes approach. Arora, Lahiri and Mukherjee (1997) relaxed the homoscedasticity assumption of Ghosh and Meeden (1986). Ghosh, Lahiri and Tiwari (1989), and Lahiri and Tiwari (1991) proposed a nonparametric empirical Bayes method that uses the Dirichlet process prior for estimating means and variances. Scott and Smith (1969) proposed a super-population model for two-stage sampling from a finite population and carried out a Bayesian predictive inference for a linear function of the finite population elements by assuming a normal prior. Their results were extended to a three-stage sampling by Malec and Sedransk (1985). Ghosh and Lahiri (1988) relaxed the normality assumption for the prior distribution and derived Bayes estimators of strata means for two-stage samples under the

assumption of posterior linearity. Meeden (1999) proposed a non-informative Bayesian approach for two-stage cluster sampling.

More references can be found in Little's work (e.g., Little 1983; 1993; 2004) and Jiang and Lahiri (2006b), which also made significant contributions to model-based/model-assisted inferences under finite population sampling.

1.8 Discussion and Layout of the Dissertation

We have given a broad review of the needs, statistical techniques, and applications of small area estimation from a historical perspective. As we reviewed, small area estimation techniques are developed not only for continuous data, but also for binary data. Many examples demonstrate that estimating proportions of units with a given characteristic for small areas by using small area estimation techniques is a common problem. When an area level model is used to produce estimates of proportions for small areas, it is commonly assumed that the survey weighted proportion for each sampled small area has a normal distribution and that the sampling variance of this proportion is known. In addition, normality is commonly assumed for the random effects of the area level or unit level mixed models. However, these assumptions need justification and may not be valid in many situations.

To tackle the above mentioned issues, in this dissertation, we develop statistical methodologies for estimating small area proportions using complex survey data based on models with better assumptions. Both area level and unit level mixed models that incorporate non-normality and non-linearity under different complex

sampling designs for dichotomous variables (say, whether a person is unemployed, or whether a baby has low birthweight) are proposed. For inference, we focus on the HB approach, although the EBP approach is also developed for one of the proposed unit level models.

This dissertation is organized as follows. In Chapter 2, we consider two new area level hierarchical Bayesian models to estimate small area proportions using survey data. To evaluate the performance of these models, we present a simulation study based on a real finite population. For each model and each dataset, the HB estimates are computed using the MCMC technique. We compare the frequentist coverage properties of these estimates with those of two existing HB models.

In Chapter 3, in order to accommodate zero direct survey estimates and the kurtosis problem for the random effects, under a one-stage sampling design we propose an adaptive HB estimation approach in which the distribution of the random effects is chosen adaptively from the exponential power class of probability distributions. The richness of the exponential power class ensures the robustness of our HB approach against departure from normality. We demonstrate the robustness of our proposed model using simulated data and several real datasets. A fully Bayesian approach based on the MCMC technique is used to make inferences.

In Chapter 4, based on the HB model proposed in Chapter 3, we study several approximate methods for Bayesian inference including first- and second-order Laplace approximations, Gauss-Hermite Quadrature, and Monte Carlo integration methods. We conduct a study using simulated data to compare the methods for the simple case when all the hyperparameters are assumed known. The results of the

study demonstrate the deficiencies of the Laplace methods. We also propose a method to conduct the analysis when all the hyperparameters are unknown.

In Chapter 5, we develop the Jiang-Lahiri type frequentist alternative to the HB methods based on the Bayesian model proposed in Chapter 3. Mean squared error formulas are developed under certain assumptions. We also propose a second-order bias corrected Taylor series linearization method and a computationally simple double parametric bootstrap method for estimating the mean squared errors of the small area estimates.

In Chapter 6, we propose a generalized linear mixed model that is suitable for binary data collected from a two-stage sampling design. The methodology developed in Chapter 3 is extended to this more complex design. Data analysis based on a sample drawn from a real finite population is conducted for evaluation purpose.

In Chapter 7, we give a summary of this dissertation and give directions for future research.

Chapter 2: Hierarchical Bayes Modeling of Survey-Weighted Small-Area Proportions

2.1 Introduction

This chapter proposes two new hierarchical Bayes models to estimate small area proportions using survey data and evaluates their performances through a Monte Carlo simulation study in which simple random samples and stratified simple random samples are generated from a fixed finite population. We compare the results obtained from these alternative models with those obtained from two commonly used models. Only HB area level modeling approach is considered in this chapter.

We organize this chapter as follows. We first review the sampling variances for direct small area proportions and discuss the problems with the current estimation methods in Section 2.2. We then introduce two commonly used models and two proposed alternative models in Section 2.3. A simulation study is presented in Section 2.4. The chapter concludes with a summary and discussion in Section 2.5.

2.2 Direct Sampling Variance and Design Effect

Let y_{ik} denote the binary response for a certain characteristic of interest for the k th unit in the i th small area $(i = 1,...,m;\ k = 1,...,N_i)$. Suppose we want to estimate the population proportion given by $P_i = \sum_{k=1}^{N_i} y_{ik} / N_i$ for the i th small area as well as the associated variance of the estimator using the sample values

y_{ik} $(i = 1,...,m;\ k = 1,...,n_i),$ drawn from the finite population under a complex sample design. We assume that $n_i > 0$ for all small areas. As we reviewed in Section 1.2, under an EPSEM design within each area, the sample proportion (mean) $p_i = \sum_{k=1}^{n_i} y_{ik} / n_i$ is an unbiased estimator of P_i. Under a NONEPSEM design within each area, a commonly used estimator is given by $p_{iw} = \sum_{k=1}^{n_i} w_{ik} y_{ik} \Big/ \sum_{k=1}^{n_i} w_{ik}$, $i = 1,\ ...,\ m$.

Let $VAR_{srs}(p_{iw})$ and $VAR_{st}(p_{iw})$ be the true variance of p_{iw} under a simple random sampling (SRS) design and a stratified SRS design within each area respectively. Following Kish (1965), the true design effect $DEFF_i$ for p_{iw} is given by

$$\begin{aligned} DEFF_i &= \frac{VAR_{st}(p_{iw})}{VAR_{srs}(p_i)} \\ &= \frac{\sum_{h=1}^{H_i} W_{ih}^2 (1 - f_{ih}) \frac{N_{ih}}{N_{ih} - 1} \frac{P_{ih}(1 - P_{ih})}{n_{ih}}}{(1 - f_i) \frac{N_i}{N_i - 1} \frac{P_i(1 - P_i)}{n_i}} \\ &\approx \frac{\sum_{\mathrm{h}=1}^{H_i} W_{ih}^2 P_{ih}(1 - P_{ih}) / n_{ih}}{P_i(1 - P_i) / n_i}, \quad \text{assuming } f_{ih} \approx 0 \text{ and } f_i \approx 0, \end{aligned} \tag{2.1}$$

where H_i is the number of strata in the i th area; n_{ih} is the sample size allocated to the h th stratum in the i th area; $n_i = \sum_{h=1}^{H_i} n_{ih}$ is the sample size for the i th area; N_{ih} is the population size of the h th stratum in the i th area; $N_i = \sum_{h=1}^{H_i} N_{ih}$ is the population size of the i th area; P_{ih} is the population proportion for the h th stratum

in the i th area; $P_i = \sum_{h=1}^{H_i} W_{ih} P_{ih}$ is the population proportion for the i th area; $f_{ih} = \frac{n_{ih}}{N_{ih}}$; $f_i = \frac{n_i}{N_i}$; $W_{ih} = \frac{N_{ih}}{N_i}$.

Equivalently, we can write $VAR_{st}(p_{iw})$ as:

$$VAR_{st}(p_{iw}) = \frac{P_i(1-P_i)}{n_i} DEFF_i. \tag{2.2}$$

Note that $DEFF_i$ is a function of P_{ih} and is unknown in practice. If $P_{ih}(1-P_{ih}) \approx P_i(1-P_i)$, $DEFF_i$ can be approximated by:

$$deff_{iw} = n_i \sum_{h=1}^{H_i} W_{ih}^2 / n_{ih}. \tag{2.3}$$

The terms in this approximation are known and can be easily obtained from the data.

Under the stratified SRS design, the direct estimator p_{iw} can be written as: $p_{iw} = \sum_{i=1}^{H_i} W_{ih} p_{ih}$, where p_{ih} is the sample proportion for the h th stratum in the i th area. The problem with p_{iw} is that it is highly unstable when the sample size n_i is small. One way to improve its precision is to borrow strength from other similar small areas. The synthetic estimation approach, which borrows strength from a large area covering the area of interest, has been considered in the literature (e.g., see Gonzales and Hoza, 1978). A synthetic estimator of P_i is given by:

$$p_w = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n_i} w_{ij} y_{ij}}{\sum_{i=1}^{m} \sum_{j=1}^{n_i} w_{ij}}, \tag{2.4}$$

the direct design-based estimator for the large area containing the i th small area. The underlying assumption of this approach, that all the small areas are very similar, is very strong and can be easily violated.

Another way to improve the precision of the small area estimates is by means of the HB approach. When an HB area level model is used to produce estimates of P_i, it is commonly assumed for the sampling model that the design-based estimates p_{iw} (or a function of them) conditioning on P_i follow normal distributions with known sampling variances. However, normality may not be a reasonable assumption if the sample sizes n_i is small or if P_i is near 0 or 1. The assumption of known sampling variances is problematic as well.

In an effort to overcome these problems, we examine two alternative models for small area proportions and compare them with two commonly used models. The models are described in the next section.

2.3 Models Studied

2.3.1 Two Commonly Used Models

We study two commonly used models for estimating small area proportions for comparison with the alternative models described in Section 2.3.3. The first is the well-known Fay-Herriot model (Fay and Herriot, 1979), which assumes known sampling variances and normal distributions for both the sampling and the linking models. The second is the normal-logistic model, which differs from the Fay-Herriot

model only by the replacement of a logit-normal distribution for the normal distribution in the linking model.

Model 1: (The Fay-Herriot model)

$$\text{Sampling model:} \quad p_{iw} \mid P_i \overset{ind}{\sim} N(P_i,\ \psi_i), \quad i = 1,...,m; \tag{2.5}$$

$$\text{Linking model:} \quad P_i \mid \boldsymbol{\beta}, \sigma_v^2 \overset{ind}{\sim} N(\mathbf{x}_i'\boldsymbol{\beta},\ \sigma_v^2), \quad i = 1,...,m. \tag{2.6}$$

Model 2: (Normal-logistic model)

$$\text{Sampling model:} \quad p_{iw} \mid P_i \overset{ind}{\sim} N(P_i,\ \psi_i), \quad i = 1,...,m; \tag{2.7}$$

$$\text{Linking model:} \quad \text{logit}(P_i) \mid \boldsymbol{\beta}, \sigma_v^2 \overset{ind}{\sim} N(\mathbf{x}_i'\boldsymbol{\beta},\ \sigma_v^2), \quad i = 1,...,m. \tag{2.8}$$

In both models the sampling variances ψ_i are assumed to be known. As we discussed in Chapter 1, Model 1 is referred to as a matched model because the sampling and linking models can be combined to produce a relatively simple linear mixed model. However, a nonlinear linking model is often preferred for modeling proportions, leading to unmatched sampling and linking models, as in Model 2 (see, for example, You and Rao, 2002a). Several link functions such as *logit*, *probit*, and *loglog* can be considered for the linking model. We choose the widely used *logit* function in order to guarantee that the estimates of P_i always fall into the right range of (0, 1).

2.3.2 Issues with Model 1 and 2

There are two main issues associated with Models 1 and 2. The first is that both models assume known sampling variances ψ_i, whereas in practice the sampling variances have to be estimated. A simple approach is to use the direct variance estimates but they are very imprecise for areas where the sample sizes n_i are small. An alternative, more complex, approach is to develop approximate estimates of P_i, say p_{isyn}, from a simple model such as a logistic model for p_{iw} in terms of the auxiliary variables, and then use these estimates in the following synthetic variance estimator:

$$\text{var}_{isyn}(p_{iw}) = \frac{p_{isyn}(1 - p_{isyn})}{n_i} deff_{iw}, \quad i = 1, ..., m. \tag{2.9}$$

In the absence of any auxiliary variable, the overall sample proportion may be used for p_{isyn} in the computation of the synthetic variance estimator (e.g., see Morris and Christiansen, 1994). The synthetic variance estimator becomes:

$$\text{var}_{isw}(p_{iw}) = \frac{p_w(1 - p_w)}{n_i} deff_{iw}, \tag{2.10}$$

where p_w is defined by (2.4). Applications of (2.9) and (2.10) need accurate, reliable estimates of the design effects.

The second issue concerns the normality assumption of the sampling model, which is based on a large sample approximation. When the sample size n_i of area i is small and P_i is near 0 or 1, as is often the case with small area estimation, the normality assumption does not work well.

2.3.3 Two Alternative Models

Under Models 1 and 2, the unknown sampling variances ψ_i are estimated in some way, and then the resultant estimates are treated as if they were known true values. An alternative approach is to treat the ψ_i as unknown parameters in the HB model. Treating the sampling variances (covariances) as unknown has been recently considered in the literature in various applications. For example, Arora and Lahiri (1997) modeled the design-based variance through the use of a HB model. Singh et al. (2005) suggested the use of a generalized design effects to smooth the sampling covariance matrix in small area modeling with survey data. More recently, You (2008) proposed the use of equal design effects over time to model the sampling variances in estimating small area unemployment rates using a cross-sectional and time series log-linear models.

We consider the following two alternative models, denoted as Models 3 and 4, that may serve to address the issues associated with Models 1 and 2. Model 3 is different from Model 2 only in the assumption made about the sampling variances ψ_i: ψ_i are assumed known in Model 2 and are assumed unknown in Model 3. The only difference between Models 3 and 4 is in the sampling distribution of the sampling model: the normal distribution is assumed in Model 3 and the beta distribution is assumed in Model 4. Model 4 was initially considered by Jiang and Lahiri (2006b) for an EBP approach in one of their illustrative examples to estimate finite population domain means.

Model 3 (Normal-logistic model with unknown sampling variance):

$$\text{Sampling model: } p_{iw} \mid P_i \overset{ind}{\sim} N(P_i,\ \psi_i), \quad i = 1,...,m; \tag{2.11}$$

$$\text{Linking model: } \text{logit}(P_i) \mid \boldsymbol{\beta}, \sigma_v^2 \overset{ind}{\sim} N(\mathbf{x}_i'\boldsymbol{\beta},\ \sigma_v^2), \quad i = 1,...,m. \tag{2.12}$$

Model 4: (Beta-logistic model with unknown sampling variance)

$$\text{Sampling model: } p_{iw} \mid P_i \overset{ind}{\sim} beta[P_i,\ \psi_i], \quad i = 1,...,m; \tag{2.13}$$

$$\text{Linking model: } \text{logit}(P_i) \mid \boldsymbol{\beta}, \sigma_v^2 \overset{ind}{\sim} N(\mathbf{x}_i'\boldsymbol{\beta},\ \sigma_v^2), \quad i = 1,...,m; \tag{2.14}$$

where $beta[\mu,\ \sigma^2]$ denotes the beta distribution with mean μ and variance σ^2. We prefer this parameterization here to the usual beta parameterization for the purpose of visual illustration of the similarity and difference between Model 4 and the other three models on the sampling model assumption. For both Model 3 and Model 4, the approximate variance function $\psi_i = [P_i(1-P_i)/n_i]deff_{iw}$ is used. Under the usual beta parameterization, $beta[P_i,\ \psi_i]$ in (2.13) becomes $beta(a_i,\ b_i)$, where the beta parameters a_i and b_i are given by:

$$a_i = P_i\left(\frac{n_i}{deff_{iw}} - 1\right), \quad b_i = (1-P_i)\left(\frac{n_i}{deff_{iw}} - 1\right).$$

The beta distribution is chosen here to model the distribution of sample proportions because: 1) it covers a rich class of distributions; 2) it may be asymmetric; 3) it is restricted to (0, 1).

HB small area estimates for all the four models can be computed using the Metropolis-Hastings algorithm within the Gibbs sampler. Details of the algorithm,

which draws random samples based on the full conditional distributions of the unknown parameters starting with one or multiple sets of initial values, are given by Robert and Casella (1999) and Chen, Shao and Ibraham (2000). You and Rao (2002a) also showed in detail how the Metropolis-Hastings algorithm within the Gibbs sampler works for models similar to Models 1 and 2. The algorithm works the same way for Models 3 and 4 as for Model 2. We include the full conditional distributions for each model in the appendix of this Chapter.

2.4 Simulation study

2.4.1 The Study Population and the Sample Designs

This section describes the simulation study that was conducted to compare the efficiency of the small area estimates produced by the four HB models. The variable of interest of our study is low birthweight. Birthweight is one of the most accessible and most understood variables in epidemiology. A baby's weight at birth is a strong indicator not only of a birth mother's health and nutritional status but also a newborn's chances for survival, growth, long-term health and psychosocial development. Babies born weighing less than 5 pounds, 8 ounces (2,500 grams) are considered as low birthweight. In contrast, the average newborn weighs about 7 pounds. Over 7 percent of all newborn babies in the United States have low birthweight. Low birthweight babies are at increased risk of serious health problems as newborns, lasting disabilities and even death. The overall birth rate of these very small babies in the

United States is increasing (*http://www.healthsystem.virginia.edu /uvahealth/peds_hrnewborn/lbw.cfm*).

There are thousands of research papers on birthweight, with hundreds more appearing every year. Wilcox (2001) compiled work from recent decades that brings a clearer understanding of what we know and do not know about birthweight. The paper summarizes why birthweight has been so popular: 1) birthweight data are free and abundant (through vital statistics); 2) birthweight is a strong predictor of an individual baby's survival (Wilcox and Russell, 1983); 3) groups with lower mean birthweight often have higher infant mortality (Wilcox, 1993; Humphrey and Elford, 1988); and 4) low birthweight is associated with poor outcomes later in life such as asthma, low IQ and hypertension (Steffensen et al., 2000; Godfrey and Barker, 2000; Nepomnyaschy and Reichman, 2006).

The simulation study was based on the 2002 Natality public-use data file. The file included all births occurring within the United States in 2002. Data were obtained from certificates filed for births occurring in each of the 50 states plus the District of Columbia (DC). We use the term "51 states" to refer to the 50 states plus DC in this study. Details about the births recorded in the National Vital Statistics System are given at the website for the National Center for Health Statistics (*http://www.cdc.gov/nchs/births.htm*).

The finite population studied comprised of 4,024,378 live birth records in the U.S. with birth weights reported. The parameters of interest are the state level low birthweight rates P_i, $i = 1,...,51$. The values of P_i varied from 5.7 percent to 11.0 percent across the states. We also computed the state level low birthweight rates by

mother's race (White, Black, and Others). Let P_i^W, P_i^B, and P_i^O denote the state level low birthweight rate among babies with White mothers only, Black mothers only, and mothers of Other race only, respectively. The values of P_i^W varied from 5.1 percent to 9.0 percent across the states, the values of P_i^B varied from 0 percent to 16.2 percent across the states, and the values of P_i^O varied from 1.9 percent to 10.8 percent across the states. Figure 2-1 displays state level low birthweight rates P_i. The x-axis of the figure represents the proportion in percentage. The 51 states were sorted by the corresponding low birthweight rates. Similarly, we display P_i^W, P_i^B, and P_i^O in Figures 2-2 to 2-4 respectively. These figures show general pictures of the true state level proportions of low birthweight babies. The state level low birthweight rates among babies with White mothers only are the smallest for most of the states among the three race groups. Except for Vermont, the state level low birthweight rates among babies with Black mothers only are at least 1.35 times larger than those with White mothers only.

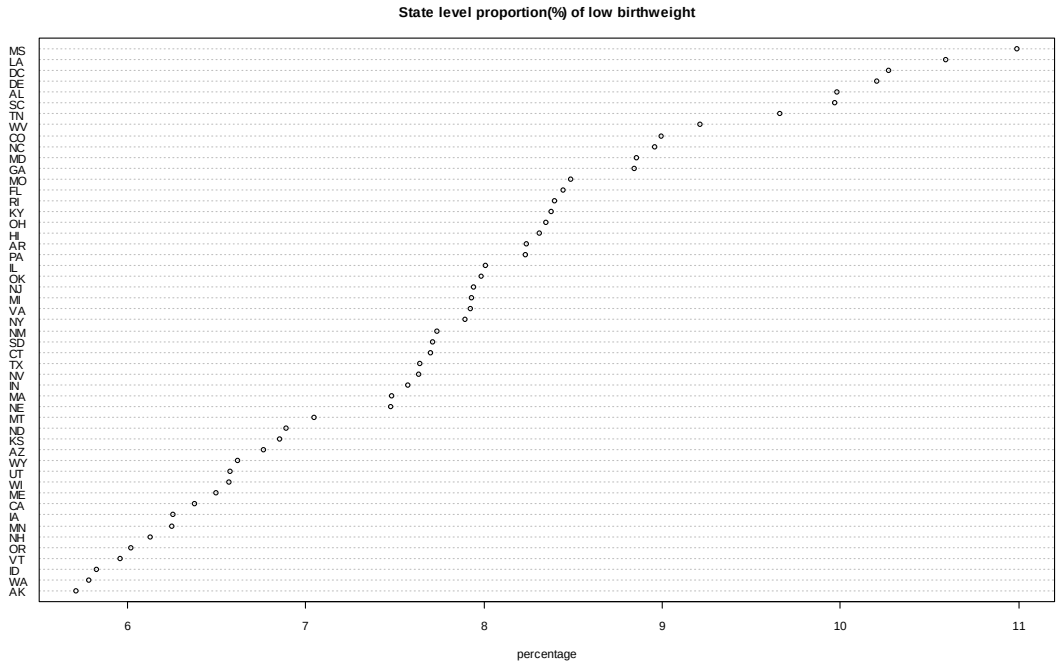

Figure 2-1: State level low birthweight rates (in percentages): P_i (states were sorted by P_i)

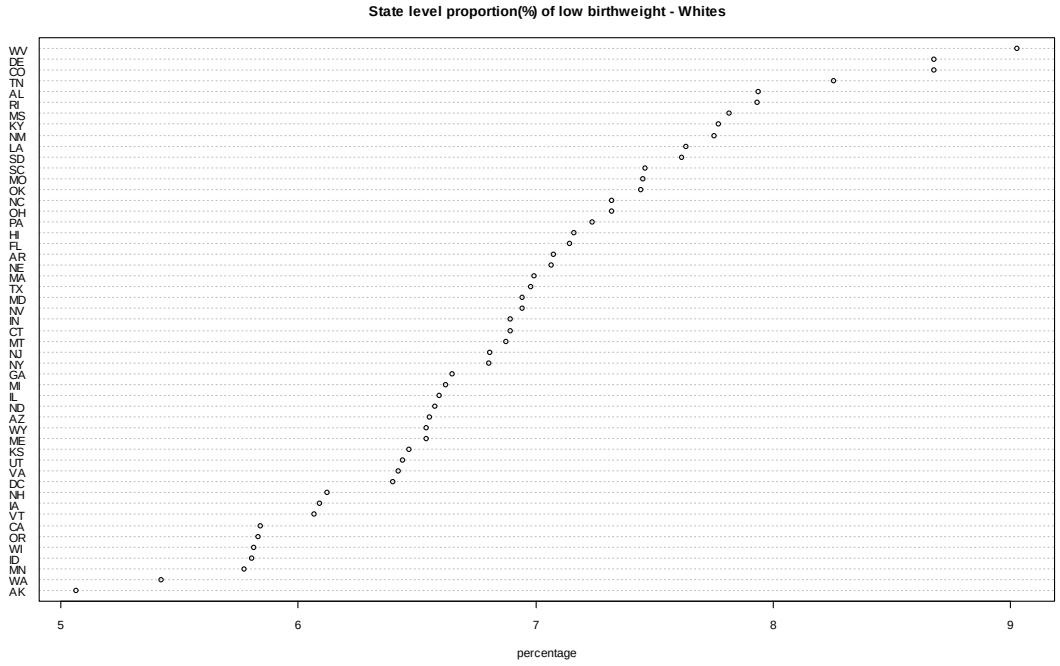

Figure 2-2: State level low birthweight rates (in percentages) among babies with White mothers only: P_i^W (states were sorted by P_i^W)

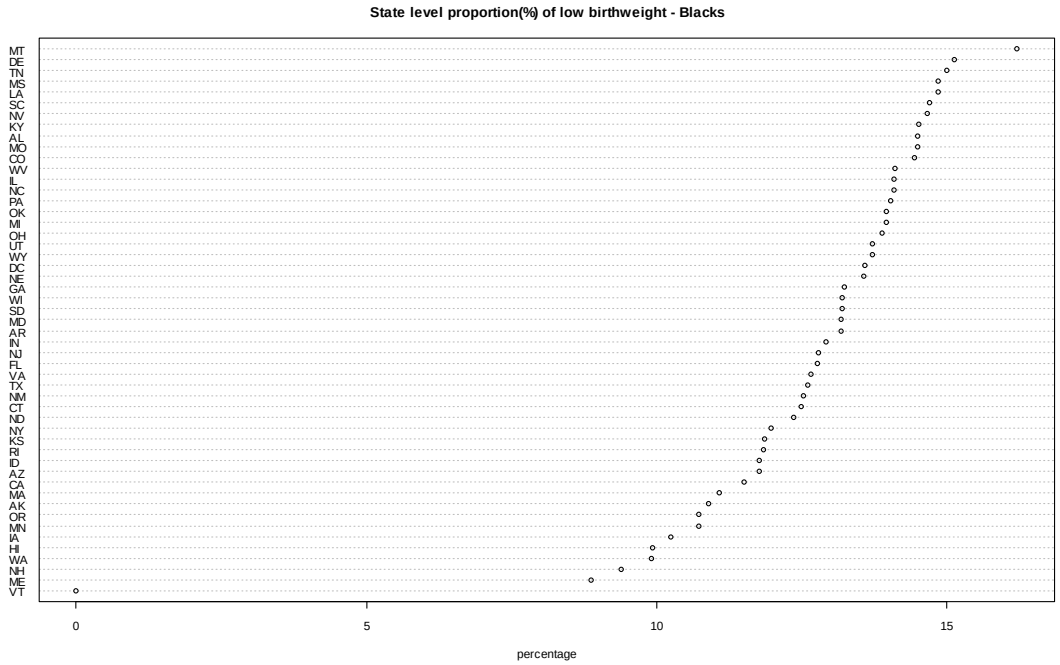

Figure 2-3: State level low birthweight rates (in percentages) among babies with Black mothers only: P_i^B (states were sorted by P_i^B)

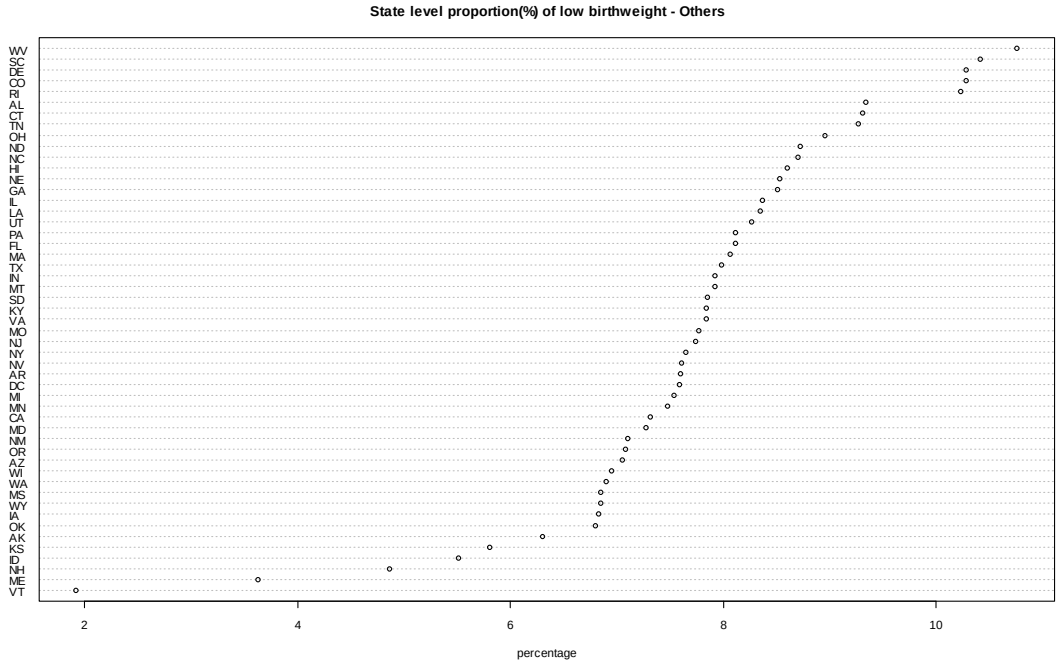

Figure 2-4: State level low birthweight rates (in percentages) among babies with mothers of Other race only: P_i^O (states were sorted by P_i^O)

Within each state, two sample designs, simple random sampling (SRS) and stratified SRS, were used to draw samples from the birth records respectively. Under the stratified SRS, mother's race was used as the stratification variable. The national sample size was set to be about 1,500 birth records for each race group in order to differentiate the sampling weights. The minority groups are oversampled based on this design. A uniform sampling fraction was used across the states for each race group subject to the condition that at least two birth records were sampled within each race group in each state. The resultant national sample size turned out to be $n = 4,526$ birth records. The state level sample sizes n_i ranged from 7 (for Vermont) to 690 (for California), with a median sample size of 61. The sample sizes n_i remained the same across different designs. This sampling procedure was repeated $R = 500$ times, creating 500 independent sample datasets under each design. Under the stratified SRS design, the sample size n_{ij} for race j in state i remained the same for each replication. Under each design, the sampling weights remained the same over different simulation runs. The sample sizes were fixed over replications in order to avoid extra random variability.

2.4.2 Auxiliary Variables

As noted in Section 1.6, auxiliary variables play an important role in small area estimation. When a large number of auxiliary variables are available, model selection techniques may be used to select a reasonable set of auxiliary variables.

We treated the finite population data as the main source for auxiliary variables because it contained many other variables in addition to the birthweight variable. Table 2-1 presents the 13 potential auxiliary variables obtained from the finite population data. Because mother's race was used as stratification variable in one design, we excluded it from the pool of potential auxiliary variables.

Table 2-1: Potential state level auxiliary variables

Index	Potential Auxiliary Variables
1	**Percentage of births with mother's age less than 15**
2	Percentage of births with mother's age less than 18
3	Percentage of births with father's age less than 15
4	**Percentage of births with father's age less than 18**
5	Percentage of births with non-Hispanic mother
6	Percentage of births with White father
7	Percentage of births with mother's education no more than high school
8	Percentage of births with native born mother
9	**Percentage of births being the first child in the family**
10	Percentage of births with no prenatal care mother
11	Percentage of births with mother whose weight gain was less than 16 pounds during pregnancy
12	Percentage of births with mother drinking alcohol during pregnancy
13	Percentage of births with mother smoking during pregnancy

Notes: 1) The percentages were computed within a state. 2) The bold variables were significant in predicting P_i at significant level $\alpha = 0.05$ based on the logistic regression model (2.15). After further HB model selection using DIC as the main criterion, we only kept variables 1 and 9 in our data analysis.

We considered DIC as the major measure to select a reasonable set of auxiliary variables from a pool of potential auxiliary variables. However, the inclusion of a large number of auxiliary variables would make the convergence of the HB models harder and hence would increase the computer running time dramatically.

In order to reduce such computational burden, we first selected a smaller set of auxiliary variables from the pool of auxiliary variables using classical methods. We then ran a set of HB models by varying the set of selected auxiliary variables, and finally applied the DIC criterion to choose the best set of auxiliary variables.

We fit the following logistic regression model using the stepwise selection procedure in SAS:

$$\text{logit}(P_i) = \beta_0 + \sum_{j=1}^{13} \beta_j x_{ij} + \varepsilon_i, \tag{2.15}$$

where P_i, $i = 1, \ldots, 51$, are the state level population proportions of low birthweight of live births obtained from the finite Natality population; x_{ij}, $j = 1,\ldots,13$, are the 13 variables listed in Table 2-1; and ε_i are the model errors that are assumed to be i.i.d. $N(0, \sigma_e^2)$. The modeling result showed that only variables 1, 4 and 9 are significant in predicting P_i at the significance level $\alpha = 0.05$, variable 6 (percentage of births with white father) is almost significant (with p-value=0.08), and all the other variables are not significant (with p-value >0.1). Starting with the three significant variables, we did a few test runs based on one sample. Based on the resulting DICs, we finally choose variable 1 (Percentage of births with mother's age less than 15) and variable 9 (Percentage of births being the first child in the family) as the final auxiliary variables to be included in our HB models.

Census 2000 data could be another source of potential auxiliary variables. For example, the state level poverty rate from Census data might be a good predictor variable for low birthweight rate. We did not consider data sources other than the

Natality database because the selection of auxiliary variables is not the main objective of this research. In addition, the inclusion of too many auxiliary variables would dramatically increase the computer running time for each HB model. As a result, for our study, we considered only the variables that were available for the finite Natality population.

2.4.3 Smoothed Sampling Variances

Both Models 1 and 2 assume that the sampling variances are known. In practice, only very imprecise estimates of the sampling variances are available. These estimates need to be "smoothed". In order to stabilize the sampling variances for the small areas, Fay and Herriot (1979) adopted the variance computations from the 1970 census process, where the sampling variances were estimated in eight states and the findings were generalized to the rest of the country. Following the same spirit, we adopted a synthetic approach to produce model-dependent estimates of the sampling variances. In each sample, we first fit the following logistic regression model on the 19 states with sample sizes $n_i > 80$ and obtained the estimate of the regression coefficient vector $(\beta_0, \beta_1, \beta_2)'$:

$$\text{logit}(p_{iw}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i, \quad i = 1,...,19, \tag{2.16}$$

where p_{iw} is the direct estimate of P_i, x_{i1} and x_{i2} are the two auxiliary variables selected in Section 2.4.2, and $\varepsilon_i \overset{iid}{\sim} N(0, \sigma_e^2)$.

We then computed a synthetic estimator of P_i for all the 51 states as follows:

$$\tilde{p}_{isyn} = \frac{\exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2}\right)}{1 + \exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2}\right)},\ i = 1,...,51, \tag{2.17}$$

where $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)'$ denotes the estimate of the coefficient vector $(\beta_0, \beta_1, \beta_2)'$ from (2.16).

We finally computed the following smoothed synthetic sampling variance of p_{iw}:

$$\tilde{v}_{isyn}(p_{iw}) = \frac{\tilde{p}_{isyn}(1 - \tilde{p}_{isyn})}{n_i} deff_{iw}, \tag{2.18}$$

where $deff_{iw}$ is defined by (2.3). The smoothed synthetic sampling variances $\tilde{v}_{syn}(p_{iw})$ defined by (2.18) were used as the final sampling variances for Models 1 and 2 and were treated as known.

For Models 3 and 4, we used sampling variance $\psi_i = [P_i(1 - P_i)/n_i] deff_{iw}$, where $deff_{iw}$ is defined by (2.3), and P_i is unknown. We estimated ψ_i concurrently with P_i through the HB modeling.

For the SRS design, $deff_{iw} \approx 1$. For the stratified SRS design, a check on the use of the approximate $deff_{iw}$ defined by (2.3) in place of $DEFF_i$ in (2.1) showed that the approximation was reasonable: the two quantities were close, with a product moment correlation of 0.96, and the ratio $deff_{iw} / DEFF_i$ varying from 0.98 to 1.30 with a mean of 1.08 and a median of 1.07.

2.4.4 Computation of the HB Estimates

The HB approach requires prior assumptions for the hyperparameters $\boldsymbol{\beta}$ and σ_v^2. The inverse-gamma prior distribution has been widely used for variance components in Bayesian data analysis in the literature (e.g., see Datta and Ghosh, 1991; You and Rao, 2002a). However, Gelman (2006) has noted that for datasets in which low values of σ_v^2 are possible, inferences under the inverse-gamma prior are very sensitive to the choice of small values of its parameters. In this sense, the non-informative uniform prior is preferable. For simplicity, we assumed the commonly used flat prior for $\boldsymbol{\beta}$, i.e., $f(\boldsymbol{\beta}) \propto 1$, and uniform prior for σ_v^2, i.e., $\sigma_v^2 \sim Uniform(0,\ L)$, where L is a large known positive number. We used $L = 100$ in this chapter. MCMC techniques as reviewed in Section 1.5.2 were used to implement the HB modeling through WinBUGS software.

For each sample dataset, the first step in the computations was to calculate the state level direct estimates. The corresponding smoothed sampling variances were also obtained following the methods described in Section 2.4.3. The direct estimates for each sample dataset were then used in turn as input to the WinBUGS software, which was used to produce the HB estimates for all four models.

For more than half of the states, the direct estimates were zero for at least one of the sample datasets. We counted the number of states with zero direct estimates (i.e., $p_{iw} = 0$) for each replicate under each sample design. Let $\phi^{(t)}$ denote the counts, $t = 1,...,500$. The values of $\phi^{(t)}$ varied from 1 to 11 among the 51 states under

the SRS design and varied from 1 to 10 under the stratified SRS design. In the six smallest states, more than 40% of the 500 direct estimates were zeros under each sample design.

Due to the zero direct estimates, WinBUGS ran into errors with undefined real results showing up for some input datasets while implementing the three HB models with the *logit* link (Models 2, 3, and 4). In order to make the Monte Carlo simulation of 500 replications run smoothly, the zero direct estimates were perturbed to very small positive numbers for the three HB models with the *logit* link.

For each WinBUGS run, three independent chains were used. For each chain, burn-ins of 10,000 samples were produced, with 10,000 samples after burn-in. The samples after burn-in were thinned to 5,000 to reduce auto-correlation of the MCMC. The resultant 15,000 MCMC samples after burn-in were then used to compute the posterior mean and percentiles for each HB model based on each sample dataset. The estimated potential scale reduction factor $\hat{R}$ proposed by Gelman and Rubin (1992) was computed for each parameter. We used $\hat{R}$ as the primary measure for convergence. The potential scale reduction factor is the factor by which the scale parameter of the estimated marginal distribution might be reduced if the simulations were continued indefinitely. The expected value of $\hat{R}$ is 1. Values of $\hat{R}$ below 1.1 are acceptable for most examples. For further details, we refer to Gelman et al. (2004, p. 296-297).

2.4.5 Simulation Results

Let P_i^{HB} denote an HB estimator of P_i, the percentage of live births with low birthweight in state i, and let $P_{i,q}^{HB}$ denote the q^{th} percentile of the posterior distribution of P_i $(i = 1,...,m)$. The noncoverage probability for the 95 percent credible intervals, i.e., the probability that the interval from $P_{i,.025}^{HB}$ to $P_{i,.975}^{HB}$ fails to cover P_i, can be estimated using the following fraction over R replications:

$$fract_i = \frac{1}{R}\sum_{r=1}^{R} I_i^{(r)},\ i = 1,...,m, \tag{2.19}$$

where $I_i^{(r)} = \begin{cases} 1, & \text{if the credible interval } \left[P_{i,.025}^{HB(r)}, P_{i,.975}^{HB(r)} \right] \text{ fails to cover } P_i \\ 0, & \text{otherwise} \end{cases}$,

$r = 1,...,R.$

The Monte Carlo simulation standard error of $fract_i$ is given by:

$$s(fract_i) = \sqrt{fract_i(1 - fract_i)/R},\ i = 1,...,m. \tag{2.20}$$

For summary purpose, the average noncoverage probability of the 95% credible intervals for an HB estimator over a group of b states, where $b \le m$, can be further estimated by:

$$Afract_b = \frac{1}{b}\sum_{i=1}^{b} fract_i. \tag{2.21}$$

The associated Monte Carlo simulation standard error of $Afract_L$ is given by:

$$s(Afract_b) = \frac{1}{b}\sqrt{\sum_{i=1}^{b}\left[s(fract_i)\right]^2}, \tag{2.22}$$

where $fract_i$ and $s(fract_i)$ are defined by (2.19) and (2.20) respectively.

The mean width (mw_i) of the credible intervals $P_{i,.975}^{HB} - P_{i,.025}^{HB}$ and its Monte Carlo simulation standard error [$s(mw_i)$] over R replications can be computed as:

$$mw_i = \frac{1}{R}\sum_{r=1}^{R}\left[P_{i,.975}^{HB(r)} - P_{i,.025}^{HB(r)} \right], \tag{2.23}$$

$$s(mw_i) = \sqrt{\frac{1}{R(R-1)}\sum_{r=1}^{R}\left[P_{i,.975}^{HB(r)} - P_{i,.025}^{HB(r)} - mw_i \right]^2}. \tag{2.24}$$

Again, for summary purpose, the average mean width (Amw_b) of the 95% credible intervals for an HB estimator over a group of b states, where $b \le m$, and the associated Monte Carlo simulation standard error [$s(Amw_b)$] over R replications can be further computed as:

$$Amw_b = \frac{1}{b}\sum_{i=1}^{b} mw_i, \tag{2.25}$$

$$s(Amw_b) = \frac{1}{b}\sqrt{\sum_{i=1}^{b}\left[s(mw_i) \right]^2}, \tag{2.26}$$

where mw_i and $s(mw_i)$ were defined by (2.23) and (2.24) respectively.

Based on results from the 500 simulation datasets for each model, Tables 2-2 and 2-3 present the following for each sample design: the noncoverage probability for the 95 percent credible intervals and the mean width of the credible intervals $P_{i,.975}^{HB} - P_{i,.025}^{HB}$ over simulations. To examine the effect of state sample size on the simulation results, the 51 states are placed into three groups according to their sample size: small $(n_i \le 30)$; medium $(30 < n_i \le 100)$; and large $(n_i > 100)$. The small group contains 16 states, the medium group contains 23 states, and the large group contains

12 states. The results presented in the tables are overall averages across all states and averages for the three groups separately.

The upper half of Table 2-2 reports the average percentage of times that the 95 percent credible interval for each P_i failed to cover the true value of P_i over the 500 replications along with the Monte Carlo simulation standard errors under the SRS design. The statistics for each model were computed based on formulas (2.21) and (2.22). The lower half of Table 2-2 displays the average widths of the 95 percent credible intervals along with the Monte Carlo simulation standard errors, which were computed based on formulas (2.25) and (2.26). The Fay-Herriot model (M1) credible intervals are the most conservative among the four models, giving about 1.5 percent overall noncoverage. The M1 credible interval widths are stable. A small proportion of the M1 credible intervals had negative lower bound. These negative lower bounds were truncated to zero. At overall 1.8 percent and 1.7 percent, the noncoverage rate of the credible intervals for the normal-logistic model (M2) and the normal-logistic model with unknown variance (M3) are very close to that for M1. Model M3 performs a little bit better in the small group compared with M1 and M2 using the noncoverage rate criterion. The noncoverage rates do not vary much across different size groups for the first three models.

At overall 10.0 percent, the noncoverage rate of the credible intervals for the beta-logistic model (M4) is well above the nominal rate of 5 percent. The noncoverage rate of 4.2 percent and 4.0 percent for the medium group and large group are closest to the nominal noncoverage rate. However, the noncoverage rate for the small group reaches 22.8 percent, pulling the overall average up. This is an

unexpected result in this study. The average width and the simulation errors are larger than those of the other three models. This instability may be due to the complexity of the full conditional distributions for the beta model. The large proportion of the 500 direct estimates that were 0 for some of the small states (see Section 2.4.4) may also cause significant problems in fitting the beta distribution. The noncoverage rates do not vary much between the medium and the large groups.

Table 2-3 displays the same results as those in Table 2-2, except that it is for stratified SRS design within each state. Table 2-3 shows a consistent pattern to that for the SRS design for the first 3 models, but with the overall and within group noncoverage rates being smaller than those shown in Table 2-2. At an overall noncoverage rate of 1.6 percent under the stratified SRS design, the performance of M3 is again showing a little bit improvement on M1 and M2 in terms of noncoverage, but it is still conservative comparing with the nominal 5 percent.

The performance of M4 improves much under the stratified SRS design. The noncoverage rate of 3.9 percent overall is closest to the nominal noncoverage rate among the four models. The noncoverage rate in the small group drops to 7.4 percent, a much more reasonable figure than 22.8 percent we saw earlier in Table 2-2. Comparing with other models under the stratified SRS design, M4 produces noncoverage rates closest to the nominal 5 percent, with reasonable average interval width, though it again has the largest simulation errors.

As expected, for all four models the mean width of the credible intervals declines with increasing state sample size. Despite these declines, however, the noncoverage rates also decline with increasing sample size for Models 3 and 4. The

noncoverage rates are in fact very small for the states with large n_i, suggesting that the credible intervals are not adequately reflecting the effect of the greater precision of the direct estimates in the states with large sample sizes.

The posterior mean of P_i based on the Fay-Herriot model M1 does not have a finite range, so it is possible that the HB estimate falling outside of the (0, 1) range. In fact, one of the HB estimates in the medium group was negative. We left it as is for model comparison purpose. In practice, negative estimates have to be fixed by alternative modeling approach. In our study, under the SRS design, 1.7 percent of the credible intervals in the small group had negative lower bounds. Under the stratified SRS design, credible intervals with negative lower bound appeared in each group. Specifically, 11.6 percent of the credible intervals in the small group, 3.9 percent of the credible intervals in the medium group, and 0.2 percent of the credible intervals in the large group had negative lower bounds.

We attempted to produce the noncoverage rates for the direct estimates as well. However, it was problematic to produce 95% confidence intervals using the regular direct estimation methods whenever the direct point estimates were zero. Alternative methods were needed to produce reasonable confidence intervals. Even for the cases with positive direct point estimates, the standard errors of the point estimates were very unstable. Therefore, we do not report the noncoverage rates for the direct estimates in this chapter.

Table 2-2: Percentage of times that the 95 percent credible intervals fail to cover P_i and mean width of the 95 percent credible intervals, along with the Monte Carlo simulation standard errors over 500 simulations (in percentages) - SRS

Sample size	M1	M2	M3	M4
Average noncoverage percentage (Monte Carlo simulation standard error)				
Overall	1.48 (0.075)	1.78 (0.082)	1.68 (0.080)	9.98 (0.175)
Small n_i	1.24 (0.123)	1.51 (0.135)	1.68 (0.142)	**22.80 (0.450)**
Medium n_i	1.68 (0.119)	1.93 (0.127)	1.83 (0.124)	4.21 (0.186)
Large n_i	1.43 (0.153)	1.87 (0.174)	1.42 (0.152)	3.97 (0.252)
Average mean width of the 95% credible intervals (Monte Carlo simulation standard error)				
Overall	6.50 (0.007)	6.51 (0.010)	6.45 (0.008)	9.25 (0.022)
Small n_i	7.52 (0.015)	7.31 (0.023)	7.13 (0.018)	10.53 (0.061)
Medium n_i	6.59 (0.010)	6.71 (0.013)	6.71 (0.012)	9.71 (0.024)
Large n_i	4.96 (0.008)	5.05 (0.010)	5.06 (0.009)	6.64 (0.014)

Notes: 1) M1 is the HB version of the Fay-Herriot model; M2 is the normal-logistic model; M3 is the normal-logistic model with unknown variance; and M4 is the beta-logistic model with unknown variance.

2) For Model 1, a small percent of the credible intervals in the small group had negative lower bounds.

3) The large noncoverage rate (22.8%) for M4 in the small group is an unexpected result. This may due to the large proportion of the 500 direct estimates that were zero for some of the small states (see section 2.4.4) and the complexity of the full conditional distributions of M4. For more explanation, see Section 2.5.

Table 2-3: Percentage of times that the 95 percent credible intervals fail to cover P_i and mean width of the 95 percent credible intervals, along with the Monte Carlo simulation standard errors over 500 simulations (in percentages) - Stratified SRS

Sample size	M1	M2	M3	M4
Noncoverage percentage (Monte Carlo simulation standard error)				
Overall	1.04 (0.063)	1.41 (0.073)	1.65 (0.079)	3.91 (0.118)
Small n_i	0.84 (0.102)	1.19 (0.121)	2.03 (0.157)	7.40 (0.283)
Medium n_i	1.14 (0.098)	1.46 (0.111)	1.53 (0.114)	2.73 (0.150)
Large n_i	1.12 (0.135)	1.60 (0.161)	1.37 (0.149)	1.53 (0.158)
Mean width of the 95% credible intervals (Monte Carlo simulation standard error)				
Overall	8.41 (0.010)	8.68 (0.018)	8.83 (0.014)	9.41 (0.019)
Small n_i	9.70 (0.023)	9.90 (0.045)	9.77 (0.034)	10.02 (0.045)
Medium n_i	8.53 (0.014)	8.86 (0.021)	9.19 (0.020)	10.00 (0.026)
Large n_i	6.48 (0.012)	6.70 (0.016)	6.89 (0.015)	7.48 (0.020)

Notes: 1) M1 is the HB version of the Fay-Herriot model; M2 is the normal-logistic model; M3 is the normal-logistic with unknown variance; and M4 is the beta-logistic model.

2) For Model 1, portion of the credible intervals with negative lower bound appeared in each group. In addition, one of the HB estimates in the medium group was negative.

In addition to the noncoverage property, we also considered the following three statistics for each estimate under each design: the overall average bias (OAB), the overall average absolute deviation (OAAD), and the overall average absolute relative deviation (OAARD) across all simulations and states which were defined as:

$$OAB = \frac{1}{mR}\sum_{i=1}^{m}\sum_{r=1}^{R}\left[p_i^{(r)} - P_i\right], \tag{2.27}$$

$$OAAD = \frac{1}{mR}\sum_{i=1}^{m}\sum_{r=1}^{R}\left|p_i^{(r)} - P_i\right|, \tag{2.28}$$

$$OAARD = \frac{1}{mR}\sum_{i=1}^{m}\sum_{r=1}^{R}\left|p_i^{(r)} - P_i\right|\Big/P_i, \tag{2.29}$$

where $p_i^{(r)}$ is the estimate of P_i based on the r th sample for state i, $m = 51,$ and $R = 500$. The corresponding Monte Carlo simulation standard errors of OAB, OAAD and OAARD were defined as:

$$s(OAB) = \frac{1}{m}\sqrt{\frac{1}{R(R-1)}\sum_{i=1}^{m}\sum_{r=1}^{R}\left[p_i^{(r)} - P_i - B_i\right]^2}, \tag{2.30}$$

$$s(OAAD) = \frac{1}{m}\sqrt{\frac{1}{R(R-1)}\sum_{i=1}^{m}\sum_{r=1}^{R}\left[\left|p_i^{(r)} - P_i\right| - AAD_i\right]^2}, \tag{2.31}$$

$$s(OAARD) = \frac{1}{m}\sqrt{\frac{1}{R(R-1)}\sum_{i=1}^{m}\sum_{r=1}^{R}\left[\frac{\left|p_i^{(r)} - P_i\right|}{P_i} - AARD_i\right]^2}, \tag{2.32}$$

where $B_i = \frac{1}{R}\sum_{r=1}^{R}\left[p_i^{(r)} - P_i\right]$, $AAD_i = \frac{1}{R}\sum_{r=1}^{R}\left|p_i^{(r)} - P_i\right|$, and $AARD_i = \frac{1}{R}\sum_{r=1}^{R}\left|p_i^{(r)} - P_i\right|\Big/P_i.$

Table 2-4 reports the OAB, OAAD, and OAARD along with the Monte Carlo simulation standard errors for each HB model under each design. The results for the direct estimates are also reported in the table. As expected, the OABs of the direct estimates under both designs were very close to zero, but the Monte Carlo simulation errors were very large. Comparing with the HB methods, the direct estimates produced the largest OAAD and OAARD with largest Monte Carlo simulation errors. This indicates that the direct estimates are highly variable. All the HB models produced negative OABs. Under both designs, the beta-logistic model (M4) produced the largest absolute OAB, OAAD and OAARD among the four HB models being considered. Those statistics for the first three HB models do not vary much. Under the stratified SRS, absolute OAB, OAAD and OAARD are increasing in the direction of M1 to M4, a direction that the model goes from simple to complex. However, such pattern does not hold under the SRS design.

Table 2-4: The overall average bias, the overall average absolute deviation, and the overall average absolute relative deviation, along with the Monte Carlo simulation standard errors over the 500 simulations and the 51 states (in percentages)

Statistics	SRS					Stratified SRS				
	M1	M2	M3	M4	Direct	M1	M2	M3	M4	Direct
OAB	-0.05	-0.09	-0.08	-0.40	0.00	-0.02	-0.10	-0.18	-0.21	0.02
$(s(OAB))$	(0.007)	(0.008)	(0.007)	(0.014)	(0.029)	(0.009)	(0.011)	(0.011)	(0.011)	(0.040)
OAAD	1.03	1.07	1.05	1.81	3.38	1.22	1.36	1.37	1.45	4.48
$(s(OAAD))$	(0.005)	(0.005)	(0.005)	(0.009)	(0.017)	(0.006)	(0.009)	(0.007)	(0.007)	(0.024)
OAARD	13.51	13.83	13.51	23.64	45.03	16.01	17.52	17.59	18.82	59.89
$(s(OAARD))$	(0.064)	(0.071)	(0.065)	(0.119)	(0.225)	(0.083)	(0.118)	(0.096)	(0.095)	(0.343)

2.4.6 Sensitivity Analysis on Model 1 and 2

To investigate how sensitive the HB models are to the prior assumptions about the variance component and the sampling variance estimation, we conducted a sensitivity analysis using Models 1 and 2. We considered two prior assumptions: 1) $\sigma_v^2 \sim Uniform(0,\ 100)$; and 2) $\sigma_v^2 \sim ING(0.001,\ 0.001)$, with the combination of two sampling variance estimators for ψ_i: i) the smoothed synthetic variance estimator defined by (2.18) in Section 2.3.3; ii) the overall synthetic variance estimator defined by (2.10) in Section 2.2.2. Thus there were four different assumptions in terms of prior and variance estimation.

Under each of the four assumptions, HB estimates were computed using WinBUGS following the set up described in Section 2.3.4. We conducted the sensitivity analysis using both sample designs for Model 1 and using the stratified SRS design for Model 2. Since the finite population is known in our study, the true variance of the sample proportion is computable. In order to evaluate the effects of the two variance estimation methods, we also carried out the HB estimates based on Model 1 using the true sampling variance and the uniform prior for σ_v^2. Table 2-5 and Table 2-6 present the summary results for each assumption by design for Model 1: the noncoverage rates for the 95 percent credible intervals, and the average of the absolute relative deviations defined by $ARD_i = \dfrac{\left|P_i^{HB} - P_i\right|}{P_i} \times 100\%$. As in Tables 2-2 and 2-3, we present the summary results by group.

Table 2-5: Percentage of times that the 95 percent credible intervals fail to cover P_i along with the Monte Carlo simulation standard errors based on 500 simulations (in percentages) for the Fay-Herriot model (M1)

	SRS					Stratified SRS				
	Uniform prior for σ_v^2			Inverse Gamma Prior for σ_v^2		Uniform prior for σ_v^2			Inverse Gamma Prior for σ_v^2	
	VAR_{st}	$\tilde{v}_{isyn}$	var_{isw}	$\tilde{v}_{isyn}$	var_{isw}	VAR_{st}	$\tilde{v}_{isyn}$	var_{isw}	$\tilde{v}_{isyn}$	var_{isw}
Overall	1.32 (0.071)	1.48 (0.075)	1.30 (0.071)	0.94 (0.060)	0.74 (0.053)	0.90 (0.059)	1.04 (0.063)	0.73 (0.053)	0.81 (0.056)	0.44 (0.041)
Small n_i	1.19 (0.120)	1.24 (0.123)	1.13 (0.117)	0.35 (0.066)	0.15 (0.043)	0.68 (0.091)	0.84 (0.102)	0.55 (0.083)	0.33 (0.064)	0.10 (0.035)
Medium n_i	1.47 (0.111)	1.68 (0.119)	1.43 (0.110)	0.91 (0.088)	0.74 (0.079)	0.83 (0.084)	1.14 (0.098)	0.75 (0.080)	0.89 (0.086)	0.37 (0.056)
Large n_i	1.20 (0.140)	1.43 (0.153)	1.28 (0.145)	1.77 (0.170)	1.52 (0.157)	1.35 (0.148)	1.12 (0.135)	0.92 (0.122)	1.32 (0.146)	1.03 (0.130)

Notes: VAR_{st} is the true variance defined by (2.2); $\tilde{v}_{isyn}$ is the smoothed synthetic variance estimator defined by (2.18); var_{isw} is the overall synthetic variance estimator defined by (2.10). The results included in the parentheses are the simulation standard errors.

The left half of Table 2-5 reports the average noncoverage rates of Model 1 under each model assumption under the SRS design, and the right half of the table reports the same statistics under the stratified SRS design. Comparing to the nominal 5 percent noncoverage rate, under the same prior assumption, the smoothed synthetic estimator $\tilde{v}_{isyn}$ defined by (2.18) appears a little bit more favorable than the overall synthetic estimator var_{isw} defined by (2.10) in terms of noncoverage. But the improvement is trivial. While under the same sampling variance estimator, the uniform prior for σ_v^2 works more favorable than the inverse-gamma prior in terms of noncoverage. Again, the improvement is not large. Under the same uniform prior

assumption for σ_v^2, the Fay-Herriot model is still conservative even if the true sampling variances were used.

Table 2-6: Absolute relative deviations along with the Monte Carlo simulation standard errors based on 500 simulations (in percentages) for the Fay-Herriot model (M1)

	SRS					Stratified SRS				
	Uniform prior for σ_v^2			Inverse Gamma Prior for σ_v^2		Uniform prior for σ_v^2			Inverse Gamma Prior for σ_v^2	
	VAR_{st}	$\tilde{v}_{isyn}$	var_{isw}	$\tilde{v}_{isyn}$	var_{isw}	VAR_{st}	$\tilde{v}_{isyn}$	var_{isw}	$\tilde{v}_{isyn}$	var_{isw}
Overall	13.06 (0.061)	13.51 (0.064)	13.12 (0.060)	15.19 (0.073)	14.71 (0.070)	15.14 (0.075)	16.01 (0.083)	14.80 (0.071)	17.50 (0.089)	16.27 (0.079)
Small n_i	15.69 (0.125)	16.14 (0.132)	15.61 (0.123)	17.44 (0.147)	16.55 (0.137)	17.86 (0.153)	19.15 (0.178)	17.56 (0.146)	20.27 (0.187)	18.50 (0.157)
Medium n_i	13.10 (0.091)	13.57 (0.095)	13.25 (0.091)	15.31 (0.110)	14.95 (0.106)	15.06 (0.112)	15.94 (0.119)	14.80 (0.106)	17.42 (0.130)	16.27 (0.119)
Large n_i	9.49 (0.093)	9.86 (0.099)	9.54 (0.094)	11.95 (0.117)	11.78 (0.114)	11.67 (0.119)	11.97 (0.122)	11.15 (0.112)	13.93 (0.138)	13.31 (0.131)

Notes: VAR_{st} is the true variance defined by (2.2); $\tilde{v}_{isyn}$ is the smoothed synthetic variance estimator defined by (2.18); var_{isw} is the overall synthetic variance estimator defined by (2.10). The results included in the parentheses are the simulation standard errors.

The left half of Table 2-6 reports the average absolute relative deviation (ARD) of Model 1 for each case under SRS design and the right half of the table presents the same statistics under the stratified SRS design. When holding the same prior for σ_v^2, the overall synthetic estimator var_{isw} produces a little bit less absolute relative deviation than the smoothed synthetic estimator $\tilde{v}_{isyn}$ does. This is expected because the overall synthetic estimator smoothes the variance extensively. On the other hand, while holding the sampling variance estimator fixed, the uniform prior for σ_v^2 leads to a little bit less absolute relative deviation than the inverse-gamma prior

does. The patterns shown in this table are consistent with those in Table 2-5. Both tables showed consistent patterns between the two designs for Model 1.

Table 2-7: Percentage of times that the 95 percent credible intervals fail to cover P_i and absolute relative deviations, along with the Monte Carlo simulation standard errors based on 500 simulations (in percentages) for the Normal-logistic model (M2) under the stratified SRS design

	Noncoverage percentage				Absolute relative deviation			
	Uniform prior for σ_v^2		Inverse Gamma prior for σ_v^2		Uniform prior for σ_v^2		Inverse Gamma prior for σ_v^2	
	$\tilde{v}_{isyn}$	var_{isw}	$\tilde{v}_{isyn}$	var_{isw}	$\tilde{v}_{isyn}$	var_{isw}	$\tilde{v}_{isyn}$	var_{isw}
Overall	1.41 (0.073)	0.88 (0.058)	3.06 (0.107)	2.71 (0.100)	17.52 (0.118)	15.78 (0.086)	15.60 (0.097)	14.35 (0.070)
Small n_i	1.19 (0.121)	0.74 (0.095)	3.91 (0.214)	3.65 (0.206)	21.06 (0.297)	18.37 (0.188)	19.26 (0.243)	17.39 (0.151)
Medium n_i	1.46 (0.111)	0.90 (0.087)	3.02 (0.158)	2.62 (0.147)	17.18 (0.142)	15.69 (0.121)	15.24 (0.119)	14.20 (0.101)
Large n_i	1.60 (0.161)	1.03 (0.130)	2.02 (0.180)	1.63 (0.163)	13.47 (0.137)	12.50 (0.126)	11.42 (0.115)	10.60 (0.102)

Notes: $\tilde{v}_{isyn}$ is the smoothed synthetic variance estimator defined by (2.18); var_{isw} is the overall synthetic variance estimator defined by (2.10). The results included in the parentheses are the simulation standard errors.

Table 2-7 presents the summary results of the noncoverage probability and the average absolute relative deviations under each assumption for Model 2 based on the stratified SRS design. The left half of Table 2-7 reports the average noncoverage rates of Model 2 under different assumptions based on the stratified SRS design, and the right half of the table reports the corresponding average absolute relative deviation. The table indicates that for Model 2, the inverse-gamma prior works better than the uniform prior in terms of both coverage property and absolute relative deviation. On the other hand, the smoothed synthetic estimator $\tilde{v}_{isyn}$ works better with respect to the

5 percent nominal noncoverage rate than the overall synthetic estimator var_{isw}, however, the latter one produces a little bit less absolute relative deviation.

In summary, the sensitivity analysis shows that for both Models 1 and 2, the smoothed synthetic estimator $\tilde{v}_{isyn}$ is superior to the overall synthetic estimator var_{isw} in terms of coverage; the uniform prior for σ_v^2 is better than the inverse-gamma prior in terms of both coverage and the absolute relative deviation for Model 1, while for Model 2, the inverse-gamma prior is superior to the uniform prior. Model 1 is less sensitive to the prior assumption than Model 2. However, many of the differences are trivial.

2.5 Summary and Discussion

In the simulation study, we have compared design-based coverage properties of credible intervals resulting from different HB models to estimate small area proportions from a simple random sample design and a stratified simple random sample design. We have also compared the HB estimates with the direct estimate. The simulation results confirmed that HB methods work better than the direct estimate in terms of both coverage and absolute relative deviation properties.

The HB version of the well-known Fay-Herriot model appears to produce the most conservative credible intervals. A big disadvantage of the Fay-Herriot model for binary data is it can produce credible intervals with negative lower bounds or even negative posterior means. The normal-logistic hierarchical model (M2) performs like the Fay-Herriot model, but its coverage is less conservative than the Fay-Herriot

model. The unknown sampling variance version of the normal-logistic hierarchical model with unknown sampling variances (M3) improves on coverage compared with the first two models, but only by a small percentage.

Compared to the other three models, the beta-logistic model with unknown sampling variances (M4) achieves credible intervals that are closest to the nominal coverage for the finite population proportions under the stratified SRS design and the proportions of the states with medium or large sample sizes under the SRS design. However, under the SRS design, this model produces credible intervals with the worst coverage among the four models for the states with small sample sizes. Model M4 did not achieve good results in terms of overall average bias, overall average absolute deviation, and overall average absolute relative deviation compared with other models. Since one of the full conditional distributions for the beta-logistic model involves the survey-weighted proportions, there is a problem with the MCMC whenever the survey-weighted proportion is zero. The credible intervals for this model are also wider than those for the other two models with a logistic linking model under the SRS design. However, the widths of the credible intervals are similar to those of the other three models under the stratified SRS design.

A second explanation for the inconsistent performance of the beta-logistic model between the two designs for the small group could be: under the SRS design, all the observations within the same area got equal sampling weights. When the state sample sizes are very small, the direct estimates p_{iw} appear more discrete than continuous; the beta distribution may be problematic to fit the discrete data. However, when the stratified SRS design is used, the observations within the same area have

different sampling weights and the variation of the weights is quite large. The unequal weighting could have improved the continuous feature of p_{iw} in the small group.

Mother's race was used as stratum variable within state under the stratified SRS design. For the concern of double counting its effect under the stratified SRS design, we did not include it in the pool of auxiliary variables when selecting auxiliary variables in Section 2.4.2. In a further study, we will include it in the pool of auxiliary variables which may help to improve the efficiency of the HB models under the SRS design.

We considered 500 simulations under each design in this study. However, the results in Tables 2-2 to 2-4 consistently showed that M4 had the largest Monte Carlo simulation errors compared with the other three HB models. A larger number of simulations are needed for M4 in order to reduce the simulation errors.

We have investigated whether different prior assumptions for the variance component σ_v^2 and different sampling variance estimation methods can affect the HB estimation results for the Fay-Herriot model and the normal-logistic model (M2). The results indicated that both the HB models are not very sensitive to the sampling variance estimation method, and M2 is more sensitive to the prior assumption for σ_v^2 than M1 is.

The simulation study was restricted to two simple sample designs. In addition, for simplicity only two auxiliary variables from one source were included in the linking models, whereas in practice the inclusion of such variables from multiple sources, especially from the census data is routine and almost essential. Further

simulation studies to cover different sample designs, different sample sizes, and to incorporate more auxiliary variables in the linking models are needed.

Appendix for Chapter 2

Appendix A: Full Conditional Distributions for the HB Models

Assume that the prior distributions for the model parameters β and σ_v^2 are $\boldsymbol{\beta} \propto 1,\ \sigma_v^2 \sim Uniform(0,\ L)$. Let $\vec{\mathbf{p}} = (p_{1w}, ..., p_{mw})'$ and $r_i = \dfrac{\psi_i}{\psi_i + \sigma_v^2}$.

The full conditional distributions for the Fay-Herriot model (M1) are given as follows:

i) $\theta_i \mid \boldsymbol{\beta}, \sigma_v^2, \vec{\mathbf{p}} \sim N\left[(1-r_i)p_{iw} + r_i \mathbf{x}_i'\boldsymbol{\beta},\ \psi_i(1-r_i)\right],$ for $\theta_i \in R$;

ii) $\boldsymbol{\beta} \mid \theta_i, \sigma_v^2, \vec{\mathbf{p}} \sim N\left[\left(\sum_{i=1}^{m} \mathbf{x}_i \mathbf{x}_i'\right)^{-1}\left(\sum_{i=1}^{m} \mathbf{x}_i \theta_i\right), \sigma_v^2\left(\sum_{i=1}^{m} \mathbf{x}_i \mathbf{x}_i'\right)^{-1}\right],$ for $\boldsymbol{\beta} \in \mathbf{R}^P$;

iii) $\sigma_v^2 \mid \boldsymbol{\beta}, \theta_i, \vec{\mathbf{p}} \sim \begin{cases} ING\left(\dfrac{1}{2}m - 1,\ \dfrac{1}{2}\sum_{i=1}^{m}(\theta_i - \mathbf{x}_i'\boldsymbol{\beta})^2\right), & \text{for } \sigma_v^2 \in (0, L) \\ 0, & \text{for } \sigma_v^2 \geq \mathrm{L} \end{cases}$.

The full conditional distributions for the normal-logistic model (M2) are given as follows:

i) $\theta_i \mid \boldsymbol{\beta}, \sigma_v^2, \vec{\mathbf{p}} \propto \dfrac{1}{\theta_i(1-\theta_i)\sigma_v\sqrt{\psi_i}} \exp\left\{-\dfrac{(p_{iw} - \theta_i)^2}{2\psi_i} - \dfrac{[\mathrm{logit}(\theta_i) - \mathbf{x}_i'\boldsymbol{\beta}]^2}{2\sigma_v^2}\right\},$

for $\theta_i \in (0,\ 1)$;

ii) $\boldsymbol{\beta} \mid \theta_i, \sigma_v^2, \vec{\mathbf{p}} \sim N\left[\left(\sum_{i=1}^{m} \mathbf{x}_i \mathbf{x}_i'\right)^{-1}\left(\sum_{i=1}^{m} \mathbf{x}_i \text{logit}(\theta_i)\right), \sigma_v^2\left(\sum_{i=1}^{m} \mathbf{x}_i \mathbf{x}_i'\right)^{-1}\right]$; for

$\boldsymbol{\beta} \in \mathbf{R}^P$;

iii) $\sigma_v^2 \mid \boldsymbol{\beta}, \theta_i, \vec{\mathbf{p}} \sim \begin{cases} ING\left(\frac{1}{2}m - 1,\ \frac{1}{2}\sum_{i=1}^{m}[\text{logit}(\theta_i) - \mathbf{x}_i'\boldsymbol{\beta}]^2\right), & \text{for } \sigma_v^2 \in (0, L) \\ 0, & \text{for } \sigma_v^2 \geq \text{L} \end{cases}$.

The full conditional distributions for the normal-logistic model with unknown variance (M3) are the same as those of M2 except that ψ_i is replaced by $\theta_i(1-\theta_i)deff_{iw}/n_i$ for the distribution of θ_i given other parameters.

Let $\delta_{iw} = \frac{n_i}{deff_{iw}} - 1$. The full conditional distributions for the beta-logistic model (M4) are given as follows:

i) $\theta_i \mid \boldsymbol{\beta}, \sigma_v^2, \vec{\mathbf{p}} \propto \frac{1}{\theta_i(1-\theta_i)\sigma_v} \frac{p_{iw}^{\theta_i\delta_{iw}-1}(1-p_{iw})^{(1-\theta_i)\delta_{iw}-1}}{\Gamma(\theta_i\delta_{iw})\Gamma[(1-\theta_i)\delta_{iw}]} \exp\left\{-\frac{[\text{logit}(\theta_i) - \mathbf{x}_i'\boldsymbol{\beta}]^2}{2\sigma_v^2}\right\}$,

for $\theta_i \in (0,\ 1)$;

ii) $\boldsymbol{\beta} \mid \theta_i, \sigma_v^2, \bar{p} \sim N\left[\left(\sum_{i=1}^{m} \mathbf{x}_i \mathbf{x}_i'\right)^{-1}\left(\sum_{i=1}^{m} \mathbf{x}_i \text{logit}(\theta_i)\right), \sigma_v^2\left(\sum_{i=1}^{m} \mathbf{x}_i \mathbf{x}_i'\right)^{-1}\right]$; for

$\boldsymbol{\beta} \in \mathbf{R}^P$;

iii) $\sigma_v^2 \mid \boldsymbol{\beta}, \theta_i, \vec{\mathbf{p}} \sim \begin{cases} ING\left(\frac{1}{2}m - 1,\ \frac{1}{2}\sum_{i=1}^{m}[\text{logit}(\theta_i) - \mathbf{x}_i'\boldsymbol{\beta}]^2\right), & \text{for } \sigma_v^2 \in (0, L) \\ 0, & \text{for } \sigma_v^2 \geq \text{L} \end{cases}$.

Appendix B: WinBUGS Code for the HB Models

WinBUGS code for Model 1:

```
model
{
for ( i in 1:N)
{
pobs[i] ~ dnorm(theta[i], D[i])
D[i] <- 1/varhat[i]
theta[i]<-inprod(beta[], X[i, ])+v[i]
v[i]~dnorm(0, tau)
}
for ( i in 1:k)
{
beta[i]~dflat()
}
tau<-1/A
A~dunif(0, 100)
}
```

WinBUGS code for Model 2:

```
model
{
for ( i in 1:N)
{
pobs[i] ~ dnorm(theta[i], D[i])
D[i] <- 1/varhat[i]
logit(theta[i])<-inprod(beta[], X[i, ])+v[i]
v[i]~dnorm(0, tau)
}
for ( i in 1:k)
{
beta[i]~dflat()
}
tau<-1/A
A~dunif(0, 100)
}
```

WinBUGS code for Model 3:

```
model
{
for ( i in 1:N)
{
pobs[i] ~ dnorm(theta[i], E[i])
E[i] <- SAMPn[i]/(theta[i]*(1-theta[i])*DEFF_kish[i])
logit(theta[i])<-inprod(beta[], X[i, ])+v[i]
v[i]~dnorm(0, tau)
```

```
D[i]<-1/E[i]
}
for ( i in 1:k)
{
beta[i]~dflat()
}
tau<-1/A
A~dunif(0, 100)
}
```

WinBUGS code for Model 4:

```
model
{
for ( i in 1:N)
{
pobs[i] ~ dbeta(a[i], b[i])
a[i] <- theta[i]*(theta[i]*(1-theta[i])/D[i]-1)
b[i] <- (1-theta[i])*(theta[i]*(1-theta[i])/D[i]-1)
logit(theta[i])<-inprod(beta[], X[i, ])+v[i]
v[i]~dnorm(0, tau)
D[i]<-theta[i]*(1-theta[i])*DEFF_kish[i]/SAMPn[i]
}
for ( i in 1:k)
{
beta[i]~dflat()
}
tau<-1/A
A~dunif(0, 100)
}
```

Chapter 3: Adaptive Hierarchical Bayesian Estimation of Small-Area Proportions

3.1 Introduction

Logistic regression mixed models typically assume normality for the area-level random effects (e.g., see model 1.11 and 1.12 in Section 1.4). The wide use of the normality assumption can be attributed to its conceptual and computational simplicity as well as its popularity in standard data analysis. Nevertheless, we would expect that certain type of measurements would not be normally distributed. For example, leptokurtic (kurtosis>0) distributions and platykurtic distributions (kurtosis<0) for individual errors can occur (e.g., see Chapter 3 of Box and Tiao, 1973). For cases where the assumption of normality is not tenable, more flexible models can be adopted to accommodate non-normality. However, the literature in small area estimation on this aspect is not rich.

Farrell, MacGibbon and Tomberlin (1994) considered the EBP approach for protecting against outlying parameters. Using a simple random-effect model which is a special case of model (1.11), Farrell et al. compared the effects of step-function priors with those of the normal and Laplace priors for the random effects. They found that as the tails of the prior become heavier, the Laplace distribution is the most appropriate prior. For skewed prior distributions, the use of a step-function prior was recommended. To the best of our knowledge, this is the only research paper addressing non-normality problem in the application of logistic regression models for estimating small-area proportions in the small area estimation literature.

In Chapter 2, we examined the performances of a normal-logistic model with unknown sampling variances (M3) and a beta-logistic hierarchical model (M4) that assumed a beta distribution for the sampling errors in the context of estimating small-area proportions. The simulation study showed that M3 works marginally better than the Fay-Herriot model in terms of noncoverage, but it is still far too conservative compared to the nominal 5 percent noncoverage. The beta-logistic model performs fine for the small areas with medium and large sample sizes in terms of coverage. However, it was problematic for handling the states with very small sample sizes due to zero survey-weighted proportions.

To accommodate zero survey-weighted proportions and non-normality related to kurtosis for the random effects, we propose robust unit level mixed models by assuming a class of distributions – the exponential power distributions, which includes the normal distribution as a special case for the random effects under complex sampling design. We make inference for small-area proportions using data from a stratified SRS design, where the small areas are the design strata. Skewness is another common feature of non-normality that may occur to many datasets. For simplicity, we do not consider it in this dissertation.

This chapter is organized as follows. We briefly review the exponential power distribution in Section 3.2. In Section 3.3, we present a motivating example for this study. In Section 3.4, we propose a robust unit level model for survey data drawn from a finite population using a stratified SRS design to accommodate kurtosis and zero problems. In Section 3.5, we illustrate some Bayesian inference procedures based on the proposed model. In Section 3.6, we evaluate the proposed model by

comparing it with the normal model using some purely simulated data and several real datasets. This chapter finishes with some concluding remarks in Section 3.7.

3.2 Exponential Power Distribution

The exponential power (EP) distribution is a three-parameter distribution whose density is given by:

$$f_{EP}(x \mid \mu, \sigma, \varphi) = \frac{c_1}{\sigma} \exp\left\{ -\left| \frac{\sqrt{c_0}}{\sigma}(x-\mu) \right|^{1/\varphi} \right\}, \; -\infty < x < +\infty$$

where $\mu \in R,\ \sigma \in \mathrm{R}^+,\ \varphi \in (0,1],\ c_0 = \Gamma(3\varphi)/\Gamma(\varphi),\ c_1 = \sqrt{c_0}\,/\left[2\varphi\Gamma(\varphi)\right]$.

The three parameters $\mu,\ \sigma,\ \varphi$ are location, scale and shape (kurtosis) parameters respectively. This parameterization is preferred to the more usual one proposed by Box and Tiao (1973) because it implies $E(X) = \mu$ and $Var(X) = \sigma^2$, a property that can be useful in modeling. This family of distributions includes a range of symmetric distributions that change gradually from the uniform ($\varphi \to 0$), through short-tailed distributions (platykurtic) to the normal ($\varphi = 0.5$), then through distributions with longer-than-normal tails (leptokurtic) to the double exponential shape ($\varphi = 1$). Figure 3-1 illustrates EP distributions with common mean $\mu = 0$ and standard deviation $\sigma = 1$ for six special values of φ. Excess kurtosis is defined as:

$$\gamma = \frac{\Gamma(\varphi)\Gamma(5\varphi)}{\Gamma^2(3\varphi)} - 3,$$

i.e., the amount of kurtosis greater (or less) than the value of 3 for a normal distribution.

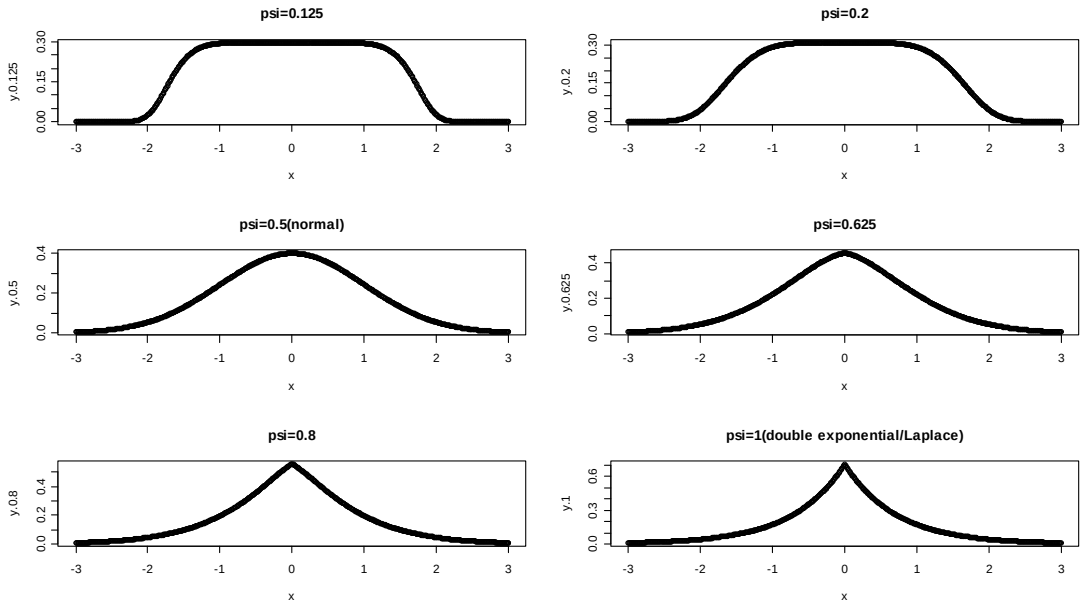

Figure 3-1: EP density plot with $\mu = 0,\ \sigma = 1$ for different φ

The exponential power distribution family can be very useful as a model in Monte Carlo robustness studies because it can attain a broad range of kurtosis values and includes three well-known symmetric distributions as special cases. Box and Tiao (1973) used this family extensively as an alternative to the normal distribution for statistical modeling and also as a tool to study Bayesian robustness. In all the examples they studied, they found that the inferences about the population mean could differ substantially as the kurtosis parameter changes. Hogg (1974) discussed the exponential power distribution family with $0.5 \le \varphi \le 1$ in relation to adaptive estimators of location. Prescott (1978) studied the asymptotic properties of the φ-trimmed means and other adaptive trimmed means from this family of distributions. For normal location problem, Choy and Smith (1997a) used the Laplace approximation method for integrals to approximate the posterior moments for the

leptokurtic class of the exponential power distribution family and found that this subclass of distributions makes the estimation procedure robust by downweighting the influence of outlying observations. For random effects models, Choy and Smith (1997b) made use of the scale mixture of normal representation of the leptokurtic density function for use in conjunction with MCMC methods.

3.3 Motivating Example – Low Birthweight Rate Data

For our evaluation purpose, we study the estimation of state level low birthweight rates using samples drawn from the 2002 Natality public-use data file. The information on the data file has been given in Section 2.4.1.

We want to fit the logistic regression model (1.11) given in Section 1.4 assuming that all the births in state i have a common probability P_i of being low birthweight. Let $(y_{ik}, x_{ik1}, x_{ik2})$ denote the indicator of low birthweight and two binary auxiliary variables (percentage of births with mother's age less than 15 and percentage of births being the first child in the family) associated with the k th baby in the i th state $(k = 1,...,N_i;\ i = 1,...,51)$, and let (P_i, x_{i1}, x_{i2}) denote the corresponding state level means. We computed P_i, x_{1i}, x_{2i} using the population data and then fitted the following logistic regression model:

$$\text{logit}(P_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + v_i, \text{ where } v_i \overset{iid}{\sim} N(0, \sigma^2),\ i = 1,...,51.$$

Both auxiliary variables are significant in predicting P_i (with p-values far less than 0.05 from the t-test). Our next goal was to assess the normality of the residuals v_i. The following methods were thereby implemented:

i) Kolmogorov-Smirnov (K-S) normality test (Stuart et al., 1999);

ii) normal Quantile-Quantile (Q-Q) plot (Gnanadesikan, 1977);

iii) a Bayesian method.

The p-value from the K-S test is 0.0436, which indicates that v_i are not normal at significance level $\alpha = 0.05$. The left panel of Figure 3-3 displays the normal Q-Q plot for v_i. The plot indicates that the underlying distribution of v_i is more like a platykurtic distribution. To verify this, we produced the descriptive statistics for v_i using SAS PROC UNIVARIATE and the results confirmed that v_i are platykurtic.

Since SAS uses a different parameterization, we consider the Bayesian approach for estimating the kurtosis of the residuals v_i. We assume a priori independence between the components of (σ, φ) and specify the following non-informative priors: i) $\varphi \sim Unif(0, 1)$ and ii) $\sigma \sim Unif(0, L)$, where L is a large positive number. As in Chapter 2, we choose $L = 100$ here. For reference on this prior assumption, we refer to Gelman (2006). We implemented model $v_i \sim EP(0, \sigma, \varphi)$ with these two prior assumptions using the WinBUGS software. The posterior mean of φ is 0.2. The one-sided 95% credible interval is (0, 0.473), which does not include the normal case ($\varphi = 0.5$). The posterior density plot of the kurtosis parameter φ is displayed on the left panel of Figure 3-2. Clearly, the posterior mode of φ is around 0.23 and the chance of covering the normal case is

very small. To assess the model fit, we applied the DIC criterion and compared the fit of the alternative models for different given kurtosis. Among several alternative models including the normal one, the EP model with $\varphi = 0.2$ fitted the data best since it resulted in the smallest DIC.

We then compared the Q-Q plot of v_i with the Q-Q plot of data randomly generated from a platykurtic exponential power distribution with $\varphi = 0.2$ (see the right panel of Figure 3-3). The similarity between the two Q-Q plots further confirms that the underlying distribution of v_i is platykurtic.

All these analyses demonstrated that a platykurtic EP distribution (with $\varphi < 0.5$) describes the underlying distribution of these residuals v_i better than the normal distribution. Based on the prior assumption $\sigma \sim Unif(0,\ 100)$, we display the posterior density plot of the scale parameter σ on the right graph of Figure 3-2. The mode of σ is around 0.12. That is, the mode of the variance σ^2 is around 0.01, which is very small. According to Gelman (2006), a uniform prior distribution is preferred here to the commonly used inverse-gamma prior distribution for the variance component σ^2 since the variance parameter σ^2 is very small.

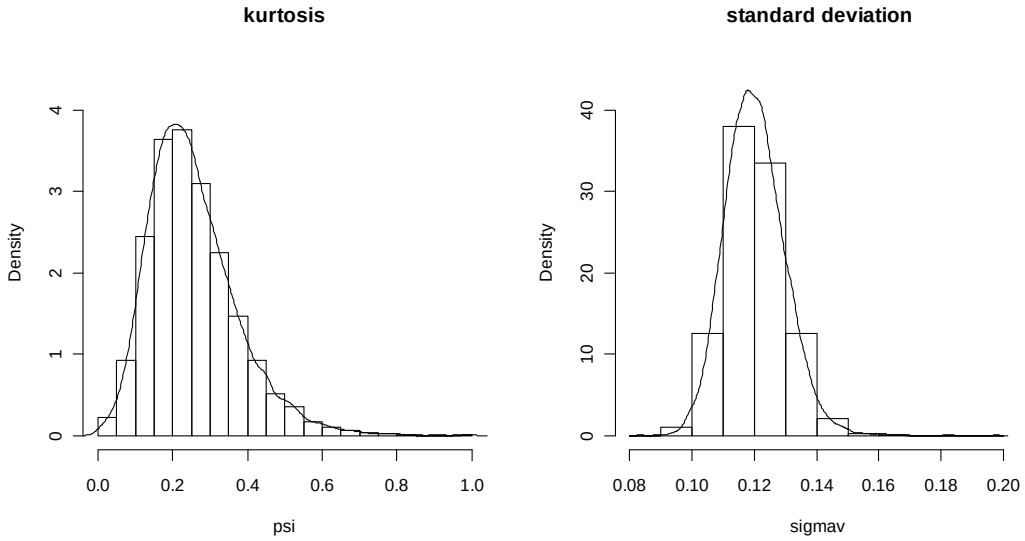

Figure 3-2: Posterior density of the hyperparameters φ and σ

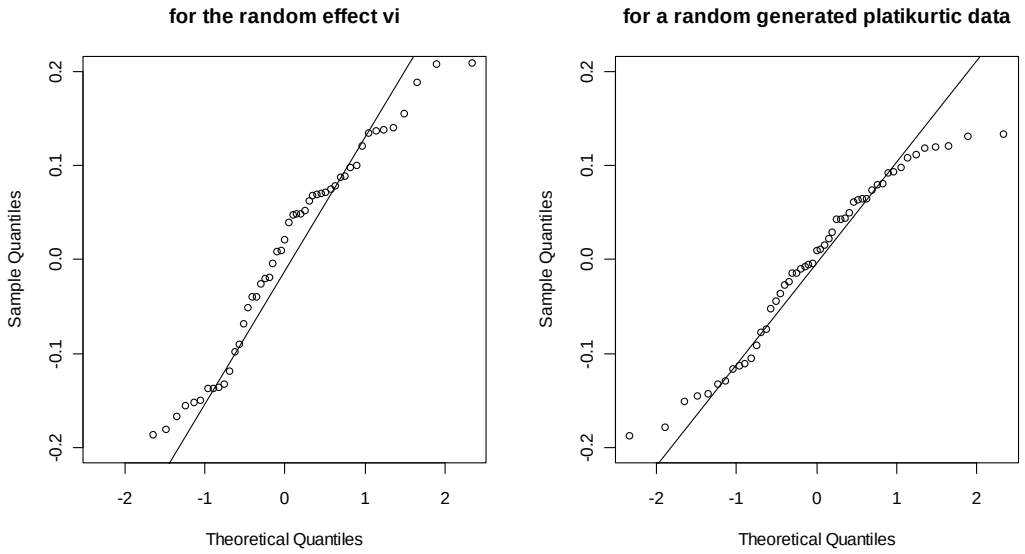

Figure 3-3: Normal Q-Q Plots for residual v_i and randomly generated data from platykurtic EP distribution

3.4 Small Area Model

Consider a finite population with m strata. Let the ith stratum be denoted by U_i with units labeled $U_{i1},...,U_{iN_i}$. Let y_{ik} denote the characteristic of interest associated with the kth unit in the ith stratum ($k=1,...,N_i;\ i=1,...,m$). Let s_i denote a random sample of fixed size n_i taken from the ith stratum using simple random sampling (SRS). Without loss of generality, suppose $s_i=(U_{i1},...,U_{in_i})$ for $i=1,...,m$, and the sample values for the characteristic of interest are denoted by $y_{i1},...,y_{in_i}$ ($i=1,...,m$). We assume no nonsampling errors are involved so that once a sample is drawn, the value of the characteristic y_{ik} is known without error. Assume that the y_{ik} are binary, that is, $y_{ik}=0$ or 1, $k=1,...,n_i;\ i=1,...,m$. Our goal is to estimate the small stratum (small area) proportions $P_i=\Sigma_{k=1}^{N_i}y_{ik}/N_i,\ i=1,\ ...,\ m$. Similar designs have been considered by other researchers (e.g., Ghosh and Meeden, 1986; 1997; Ghosh and Lahiri, 1987a, b; Nandram and Sedransk, 1993; Jiang and Lahiri, 2001; MacGibbon and Tomberlin, 1989).

Under this design, in order to estimate the small area proportions P_i, $i=1,...,m,$ the following basic logistic mixed effect model is commonly used (e.g., see Jiang and Lahiri, 2006b):

$$\text{Level 1: } y_{ik}\,|\,\theta_i \overset{ind}{\sim} \text{Bernoulli}(\theta_i);\qquad k=1,...,n_i;\ i=1,...m, \tag{3.1}$$

$$\text{Level 2: } \text{logit}(\theta_i)=\mathbf{x}_i'\boldsymbol{\beta}+v_i\ , \tag{3.2}$$

$$\text{where } v_i \overset{iid}{\sim} N(0,\sigma_v^2),\ i=1,...,m\,. \tag{3.3}$$

Here θ_i, $i = 1,...,m$ are the model parameters for the expectation of y_{ik}, $k = 1,...,n_i;\ i = 1,...,m$. For convenience, we call the model (3.1)~(3.3) the *Bernoulli-Logit-Normal* model. As demonstrated in Section 3.3, the normal assumption for the random effects v_i in (3.3) does not allow for the possibility of kurtosis. One can possibly assume a specific distribution from the exponential power family such as Laplace (double exponential) distribution. However, there is still a mis-specification risk for the distribution of the random effects. To improve robustness, instead of assuming the normal or some other specific non-normal distribution, we assume that the random effects v_i follow an unspecified distribution belonging to the exponential power distribution family with two parameters:

$$v_i \overset{iid}{\sim} EP(0, \sigma_v, \varphi_v). \tag{3.4}$$

We call the proposed model (3.1)-(3.2)-(3.4) the *Bernoulli-Logit-EP* model. We assume the hyperparameters σ_v and φ_v are both unknown. The strength of this model is that we use a class of probability distributions instead of a specific one and the underlying model will be determined by the data.

The EP density has been considered by Fabrizi and Trivisano (2007) as one of their robust extensions to the Fay-Herriot model for continuous data. The idea of using a class of distributions instead of a specific one for model-based inference of finite population total can also be found in Li and Lahiri (2007), where a super-population model was chosen adaptively from the well-known Box-Cox class of transformations. However, they did not consider a small area application, which is more complex because of the presence of random effects.

3.5 Bayesian Inference

We are interested in estimating the finite small-area proportions P_i, $i = 1, \ldots, m$, based on the *Bernoulli-Logit-EP* model. Let s_i and s_i^c denote the set of sampled units and non-sampled units respectively, and let $\mathbf{y}_s = \{y_{11}, \ldots, y_{1n_1}, \ldots, y_{m1}, \ldots, y_{mn_m}\}'$. The Bayes estimator of P_i is the mean of the posterior distribution of P_i. We can write P_i as:

$$\begin{aligned} P_i &= \frac{1}{N_i}\left(\Sigma_{k\in s_i} y_{ik} + \Sigma_{k\in s_i^c} y_{ik}\right) \\ &= \frac{1}{N_i}\left[n_i p_i + \left(N_i - n_i\right) p_{ins}\right] \\ &= f_i p_i + (1 - f_i) p_{ins}, \end{aligned} \tag{3.5}$$

where $f_i = n_i / N_i$ are the sampling rates and $1 - f_i$ are the finite population corrections, $p_i = \sum_{k\in s_i} y_{ik} \Big/ n_i$ are the area level proportions based on the sample units only, and $p_{ins} = \sum_{k\notin s_i} y_{ik} \Big/ \left(N_i - n_i\right)$ are the area level proportions based on the nonsampled units only. Since the p_i are known given the sample, from (3.5), we can say that the prediction of P_i is equivalent to the prediction of p_{ins} given the sample.

The Bayes estimator of P_i is:

$$
\begin{aligned}
E(P_i \mid \mathbf{y}_s) &= f_i p_i + (1-f_i) E\left(p_{ins} \mid \mathbf{y}_s\right) \\
&= f_i p_i + (1-f_i) \frac{1}{N_i - n_i} \sum\nolimits_{k \in s_i^c} E\left(y_{ik} \mid \mathbf{y}_s\right) \\
&= f_i p_i + (1-f_i) \frac{1}{N_i - n_i} \sum\nolimits_{k \in s_i^c} E\left[E\left(y_{ik} \mid \theta_i, \mathbf{y}_s\right) \mid \mathbf{y}_s\right] \\
&= f_i p_i + (1-f_i) E\left(\theta_i \mid \mathbf{y}_s\right) \equiv g_i\left[E(\theta_i \mid \mathbf{y}_s)\right],
\end{aligned}
\tag{3.6}
$$

where $\theta_i = \dfrac{\exp(\mathbf{x}_i'\boldsymbol{\beta} + v_i)}{1+\exp(\mathbf{x}_i'\boldsymbol{\beta} + v_i)}$. From (3.6), we can see that once $E(\theta_i \mid \mathbf{y}_s)$ is estimated, it is straightforward to estimate $E(P_i \mid \mathbf{y}_s)$ if f_i is known. We can also see that $E\left(\theta_i \mid \mathbf{y}_s\right)$ is a good approximation of $E(P_i \mid \mathbf{y}_s)$ if $f_i \approx 0$.

Further, the posterior variance of P_i is given by:

$$
V(P_i \mid \mathbf{y}_s) = V\left[E(P_i \mid \mathbf{y}_s, \theta_i) \mid \mathbf{y}_s\right] + E\left[V(P_i \mid \mathbf{y}_s, \theta_i) \mid \mathbf{y}_s\right]. \tag{3.7}
$$

Since $E(P_i \mid \mathbf{y}_s, \theta_i) = f_i p_i + (1-f_i)\theta_i$, and $V(P_i \mid \mathbf{y}_s, \theta_i) = \dfrac{1}{N_i}(1-f_i)\theta_i(1-\theta_i)$, formula (3.7) can be further written as:

$$
\begin{aligned}
V(P_i \mid \mathbf{y}_s) &= (1-f_i)^2 V(\theta_i \mid \mathbf{y}_s) + \frac{1}{N_i}(1-f_i) E\left[\theta_i(1-\theta_i) \mid \mathbf{y}_s\right] \\
&= (1-f_i)^2 V(\theta_i \mid \mathbf{y}_s) + \frac{1}{N_i}(1-f_i)\left[E(\theta_i \mid \mathbf{y}_s) - V(\theta_i \mid \mathbf{y}_s) - E^2(\theta_i \mid \mathbf{y}_s)\right] \\
&= (1-f_i)\left\{\left(1 - f_i - \frac{1}{N_i}\right) V(\theta_i \mid \mathbf{y}_s) + \frac{1}{N_i} E(\theta_i \mid \mathbf{y}_s)\left[1 - E(\theta_i \mid \mathbf{y}_s)\right]\right\} \\
&\equiv h_i\left[V(\theta_i \mid \mathbf{y}_s),\ E(\theta_i \mid \mathbf{y}_s)\right],
\end{aligned}
\tag{3.8}
$$

Formula (3.8) indicates that $V(P_i \mid \mathbf{y}_s)$ is a linear function of $V(\theta_i \mid \mathbf{y}_s)$ and $E(\theta_i \mid \mathbf{y}_s)$. Once $V(\theta_i \mid \mathbf{y}_s)$ and $E(\theta_i \mid \mathbf{y}_s)$ are obtained, it is straightforward to

obtain $V(P_i \mid \mathbf{y}_s)$ if f_i and N_i are known. We can also see from (3.8) that $V\left(P_i \mid \mathbf{y}_s\right) \approx V\left(\theta_i \mid \mathbf{y}_s\right)$ if $f_i \approx 0$.

Once $E(P_i \mid \mathbf{y}_s)$ and $V(P_i \mid \mathbf{y}_s)$ are obtained, the posterior density of P_i can be approximated by the normal density with mean $E(P_i \mid \mathbf{y}_s)$ and $V(P_i \mid \mathbf{y}_s)$. That is:

$$P_i \mid \mathbf{y}_s \overset{ind}{\sim} N\left[E(P_i \mid \mathbf{y}_s),\ V(P_i \mid \mathbf{y}_s)\right]. \tag{3.9}$$

It is easy to make any inference on P_i such as posterior mean, posterior variance, credible intervals, using the posterior density of P_i.

In this chapter, we consider the case when $f_i \approx 0$ which occurs often when there are many small areas. As we demonstrated earlier, the inference on P_i is equivalent to the inference on θ_i if $f_i \approx 0$. As a result, the Bayesian inference will be focused on the posterior distribution:

$$f(\theta_1,...,\theta_m \mid \mathbf{y}_s) = \int_{\boldsymbol{\beta}} \int_{\sigma_v} \int_{\varphi_v} f(\theta_1,...,\theta_m, \boldsymbol{\beta}, \sigma_v, \varphi_v \mid \mathbf{y}_s) d\boldsymbol{\beta} d\sigma_v d\varphi_v .$$

Based on the small area models described in Section 3.4, the joint posterior distribution $f(\theta_1,...,\theta_m, \boldsymbol{\beta}, \sigma_v, \varphi_v \mid \mathbf{y}_s)$ cannot be expressed in a single closed form; hence an approximation is needed. However, the joint posterior distribution can be simulated using a MCMC method, such as Gibbs sampling or the Metropolis-Hastings algorithm. Following Malec et al. (1997), we will make inference about θ_i through HB approach and implement the proposed model using the MCMC technique. The posterior mean $E\left(\theta_i \mid \mathbf{y}_s\right)$ approximates the HB point estimate of P_i and the posterior variance of $V\left(\theta_i \mid \mathbf{y}_s\right)$ is used as a measure of variability.

The HB approach requires prior assumptions about the hyperparameters $\boldsymbol{\beta}$, σ_v, and φ_v. Assume they are independent, that is, $f(\boldsymbol{\beta}, \sigma_v, \varphi_v) = f(\boldsymbol{\beta}) f(\sigma_v) f(\varphi_v)$. We draw samples $\{\theta_1^{(d)}, ..., \theta_m^{(d)}, \boldsymbol{\beta}^{(d)}, \sigma_v^{(d)}, \varphi_v^{(d)}; d = 1, ..., T\}$ from the joint posterior distributions $f(\theta_1, ..., \theta_m, \boldsymbol{\beta}, \sigma_v, \varphi_v \mid y_s)$ using the Metropolis-Hastings algorithm within the Gibbs sampler. Details of the algorithm, which draws random samples based on the full conditional distributions of the unknown parameters starting with one or multiple sets of initial values, are given by Robert and Casella (1999) and Chen, Shao, and Ibraham (2000).

3.6 Model Evaluation and Data Analysis

In this section, we evaluate the robustness of our proposed model using both simulated data and real data.

3.6.1 Simulated Data Analysis

The aim is to compare the *Bernoulli-Logit-EP* and *Bernoulli-Logit-Normal* models with the random effects v_i generated under different distributions. In this simulation exercise, we would like to investigate the following issues:

1) When v_i are non-normal, whether the *Bernoulli-Logit-EP* model is more effective than the *Bernoulli-Logit-Normal* model;
2) When v_i are actually normal, what is the effect of overparameterization by the *Bernoulli-Logit-EP* model.

To generate the data, we set $n_i = 5$ and $m = 100$. We also set four different cases of σ and φ by varying the values:

i) $\sigma_v^2 = 0.01$ and 0.1;

ii) $\varphi_v = 0.2$ (Platykurtic) and 0.5 (Normal).

For each of the four combinations, we generated one sample dataset from the models: $\text{logit}(\theta_i) = \mu + v_i$, $v_i \sim EP(0, \sigma_v, \varphi_v)$, $i = 1, ..., m$ and $y_{ij} \sim Bernoulli(P_i)$, $j = 1, ..., n_i$, $i = 1, ..., m$. Without loss of generality, we set $\mu = 0$.

To implement the HB modeling using the sampled data, we assume that no auxiliary variables are available, i.e., $\mathbf{x}_i'\boldsymbol{\beta} = \mu$; we also specify the following prior assumptions for individual parameters: i) Flat prior for μ, i.e., $f(\mu) \propto 1$; ii) $\sigma_v \sim Uniform(0,\ L)$, and iii) $\varphi_v \sim Uniform(0,\ 1)$.

Using the data from each sample as input, we computed HB estimates for the two models using WinBUGS. For each WinBUGS run, three independent chains were used. For each chain, burn-ins of 1,000 samples were produced, with 4,000 samples after burn-in. The resultant 12,000 MCMC samples after burn-in were then used to compute the posterior mean and percentiles for each HB model based on each sample dataset. The potential scale reduction factor $\hat{R}$ was used as the primary measure for convergence (see Gelman and Rubin, 1992).

Let θ_i^{HB} denote an HB estimator of θ_i, and let $\theta_{i,q}^{HB}$ denote the q^{th} percentile of the posterior distribution of θ_i. To evaluate the two HB models, the following two evaluation statistics for each HB estimator are calculated:

- Average absolute deviation (AAD), $AAD = \frac{1}{m}\sum_{i=1}^{m}\left|\theta_i^{HB} - \theta_i\right|$
- Average absolute relative deviation (AARD),

$$AARD = \frac{1}{m}\sum_{i=1}^{m}\left|\theta_i^{HB} - \theta_i\right| \Big/ \theta_i$$

Table 3-1 reports the ratios of AAD and AARD for the HB estimates based on the model *Bernoulli-Logit-Normal* over those based on the alternative model. When the random effects v_i were generated from the EP distribution with $\varphi = 0.2$ (see the first two rows in the table), the *Bernoulli-Logit-Normal* model gives poorer results than the *Bernoulli-Logit-EP* model. For example, the loss is over 13 percent in terms of both SRASRD and AARD for the first case. When the random effects v_i were generated from normal distributions (see the last two rows in the table), the *Bernoulli-Logit-EP* model gives fair results compared with the *Bernoulli-Logit-Normal* model. Overall, the results indicate that the effect of overparameterization is not worrisome in this example and that the *Bernoulli-Logit-EP* model is robust. The table also shows that when σ_v^2 is larger, the results from the two models are closer, which means that the results are less sensitive to the kurtosis measure.

Table 3-1: Ratios of AAD and AARD for the two models (Normal/EP) using simulated data

Data generating distribution	AAD	AARD
$EP(\mu = 0,\ \sigma_v = 0.1,\ \varphi_v = 0.2)$	1.131	1.132
$EP(\mu = 0,\ \sigma_v = 0.33,\ \varphi_v = 0.2)$	1.029	1.027
$N(\mu = 0,\ \sigma_v^2 = 0.01)$	0.992	0.992
$N(\mu = 0,\ \sigma_v^2 = 0.11)$	0.996	0.996

3.6.2 Real Data Analysis

In this subsection, we first conduct data analysis using samples drawn from a real finite population, the 2002 Natality public-use data file. We then conduct the analysis based on two real datasets: the well-known baseball data (Efron and Morris 1975) and the 1994 Missouri turkey hunting data (He and Sun, 1998).

1. Sample Data Drawn from the 2002 Natality Population

We revisit the birthweight problem using the 2002 Natality public-use data as described in Section 3.3. Instead of running a simulation, we drew $R = 6$ sets of independent samples of size $n = 4,526$ using simple random sampling within states from the finite population. The state level sample sizes n_i ranged from 7 (for small states such as Vermont) to 690 (for California). The sample sizes are the same as the ones used in Chapter 2. The sampling fractions f_i varied from 0.0007 to 0.0046 which are approximatly equal to zero, so the *fpc* can be ignored. We did not consider the stratified SRS design for this analysis because it requires more complex models than the models studied in this chapter.

In this analysis, we want to compare the performance of the two models: *Bernoulli-Logit-EP* and *Bernoulli-Logit-Normal.* In addition, in order to further motivate the preference of the proposed unit-level mixed model over the area level models studied in Chapter 2 for our problem, we would like to compare the performance of the proposed *Bernoulli-Logit-EP* model with the two area level

models, normal-logistic model with unknown variance (M3) and beta-logistic model (M4) , studied in Chapter 2 as well.

Using data from each sample, we computed the HB estimates for all the four models incorporating the two auxiliary variables used in Section 3.3. The prior distributions on the hyperparameters are identical to the ones used in Section 3.6.1.

To compare the four HB models, the two evaluation statistics, described in Section 3.6.1, are again computed for each HB estimator. Table 3-2 reports the ratios of the evaluation statistics for *Bernoulli-Logit-Normal*, M3, and M4 to B*ernoulli-Logit-EP*. We do not report the simulation errors in this table because we only considered six replicate samples. The numbers in the table consistently show that the *Bernoulli-Logit-EP* model works better than the *Bernoulli-Logit-Normal* model in terms of the four evaluation statistics. The table also shows that the two unit level models perform better than the two area level models. The performance of the beta-logistic model (M4) is not good especially with the sixth sample. This is consistent with what we obtained in Chapter 2 under the SRS design. As we demonstrated in Section 3-3, the random effects v_i for this dataset are not normally distributed. Therefore, the analysis result in this subsection is consistent with what we found using purely simulated data in Section 3.6.1.

Table 3-2: Ratio of the two summary statistics for three HB estimators over those for the HB estimator based on the *Bernoulli-Logit-EP* model using the Natality data

Sample	Model	AAD	AARD
1	Bernoulli-Logit-Normal	1.016	1.016
1	normal-logistic (M3)	1.037	1.030
1	beta-logistic (M4)	1.477	1.519
2	Bernoulli-Logit-Normal	1.045	1.040
2	normal-logistic (M3)	1.103	1.078
2	beta-logistic (M4)	1.320	1.254
3	Bernoulli-Logit-Normal	1.021	1.012
3	normal-logistic (M3)	1.142	1.130
3	beta-logistic (M4)	1.879	1.973
4	Bernoulli-Logit-Normal	1.023	1.016
4	normal-logistic (M3)	1.018	1.015
4	beta-logistic (M4)	1.719	1.713
5	Bernoulli-Logit-Normal	1.076	1.076
5	normal-logistic (M3)	1.153	1.147
5	beta-logistic (M4)	1.525	1.496
6	Bernoulli-Logit-Normal	1.014	1.013
6	normal-logistic (M3)	1.037	1.028
6	beta-logistic (M4)	2.396	2.382

Note: The denominators of the ratios are the corresponding estimates from the Bernoulli-Logit-EP model.

2. Baseball Data

In this subsection, we revisit the well-known baseball data given in Efron and Morris (1975). This dataset has been analyzed by several researchers in the past, including Efron and Morris (1975), Morris (1983), Gelman et al. (2004), Datta and Lahiri (2000), Rao (2003), Jiang and Lahiri (2006a), among others. The dataset contains the batting averages of 18 major league players through their first 45 official at bats of the 1970 season (p_i) and the true batting averages of all the 18 players for the rest of the 1970 season (p_{ins}). Efron and Morris (1975) used this dataset to

demonstrate the performance of their empirical Bayes and limited translation empirical Bayes estimators derived using an exchangeable prior in the presence of an outlying observation. They considered the problem of predicting the batting average for all the players for the remainder of the 1970 season based on their batting averages for the first 45 at bats. Gelman et al. (2004) provided additional data for this estimation problem and included important auxiliary data like the batting average of each player in the previous (1969) season. We consider the same estimation problem as Efron and Morris (1975) did. That is, we want to predict p_{ins} using the sampled data.

The sample size $n_i = 45$ is the number of times at bats for each player, $i = 1,...,18$. We computed the HB estimates for p_{ins} using the two models based on the baseball data: *Bernoulli-Logit-EP* and *Bernoulli-Logit-Normal*. The previous season batting average was used as a covariate. From (3.6) and (3.8), we can derive that $E(p_{ins} \mid y_s) = E(\theta_i \mid y_s)$ and $V(p_{ins} \mid y_s) = V(\theta_i \mid y_s)$.

For each player, Figure 3-4 displays the true batting average for the rest of the 1970 season (Ptrue) along with the sample proportion (DirectP) and the two different HB estimators (HBEP and HBNorm) in the increasing order of the previous season average. The figure shows that two HB estimates are very close to each other and performed much better than the direct estimates. Table 3-3 reports the two summary statistics for both models and it further confirms the closeness of the two HB estimators.

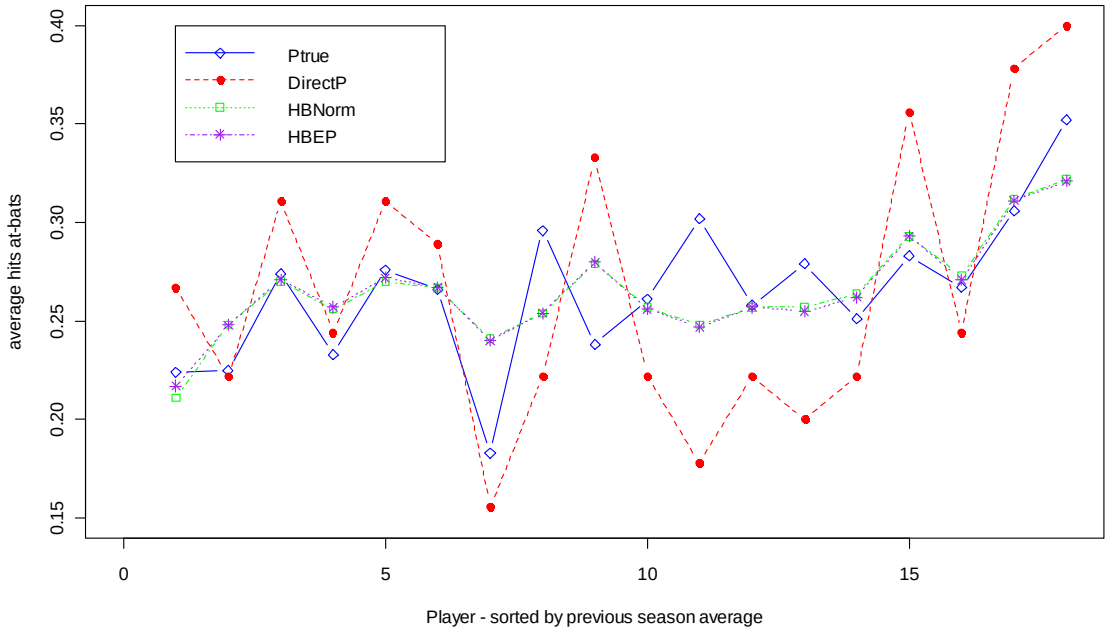

Figure 3-4: HB estimates of the batting averages for the rest of the 1970 season

Table 3-3: Summary statistics for the two HB estimators using the baseball data

Model	AAD	AARD
Bernoulli-Logit-EP	0.0195	0.077
Bernoulli-Logit-Normal	0.0198	0.079

The true values of p_{ins} are available for the baseball data. In order to investigate the nature of the random effects v_i, we fitted the logistic regression model considered in Section 3.3 on p_{ins} incorporating the previous season average at bats as the covariate. We then tested the normality of the residuals v_i using the Kolmogorov-Smirnov (K-S) normality test and the normal Q-Q plot. The $p-$value of 0.676 from the K-S test concludes that v_i appear to be normal. Figure 3-5 displays the normal Q-Q plot of the residuals v_i. One player on the extreme left of the graph appears as

an outlier. Excluding that outlier, v_i look approximately normal. The posterior mean of the kurtosis parameter φ_v estimated using the *Bernoulli-Logit-EP* model equals to 0.506, further confirming the approximate normality of v_i .

The finding from this analysis is consistent with the simulated data analysis, that is, when the random effects v_i are actually normal, the over-parameterization is not worrisome.

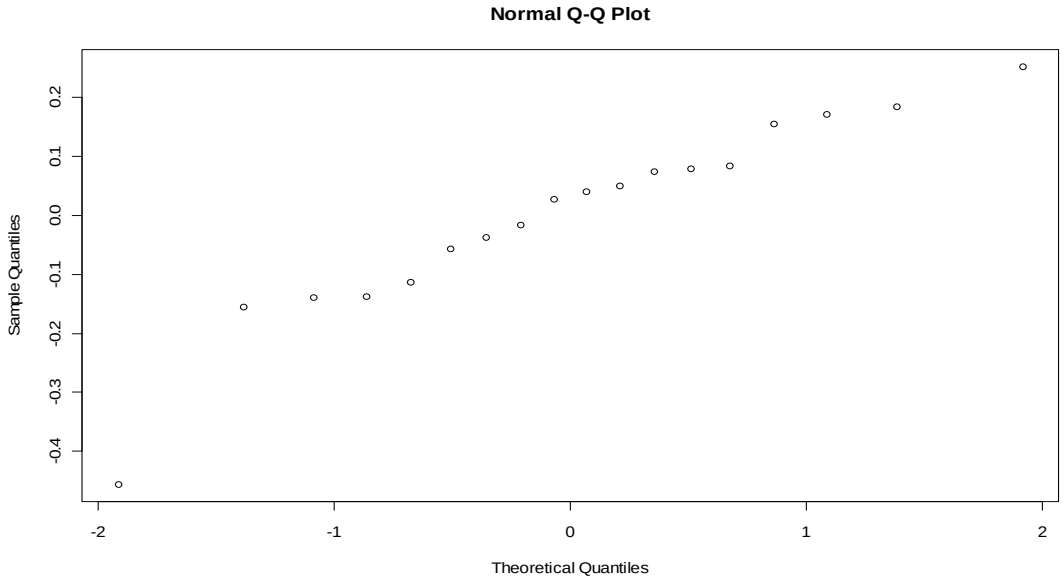

Figure 3-5: Q-Q plot of the residuals v_i based on Baseball data

3. Missouri Turkey Hunting Survey Data

The Missouri Turkey Hunting Survey (MTHS) is a bi-annual postseason mail survey conducted by the Missouri Department of Conservation to monitor and aid in the regulation of the turkey hunting season. Questionnaires are mailed to a random sample of permit buyers after the turkey hunting season. The MTHS provides

information concerning the number of turkeys harvested by hunters on each day of the hunting season and the total number of trips made to the counties by these hunters on each hunting day. The success rates are then obtained from this information. The 1994 spring season data was analyzed by He and Sun (1998). Let n_i be the total number of trips made by the sampled hunters to county i, y_i be the number of successful trips among the sample of n_i, and P_i be the probability of success for each trip in county i. The problem was to estimate the county specific success rates θ_i for all counties in Missouri. He and Sun (1998) provided HB estimates of success rates for all the 114 counties in Missouri using a simple Binomial-Beta model without covariates. With the 1996 spring season data, He and Sun (2000) estimated the county level success rates using a hierarchical Bayesian generalized linear model with spatial correlations.

We revisit the 1994 spring season data analyzed by He and Sun (1998) in this subsection. We excluded three counties with zero sample size from our data analysis for simplicity. They can be predicted from the same model using the parameters estimated from the rest of the data. The sample sizes for the other 111 counties varied from 2 to 802. The total numbers of trips N_i made by the population of hunters were unknown, so we assumed the sampling fractions $f_i \approx 0$. In addition, there were no covariates available for the data analysis. We computed the HB estimates for the 111 counties using the models *Bernoulli-Logit-EP* and *Bernoulli-Logit-Normal* under the same prior assumptions considered in the earlier sections.

Figure 3-6a displays the two HB estimates (HBEP and HBNorm) along with the direct sample estimates (DirectP) sorted by the sample size in the increasing

order. Since we do not know the true proportions P_i for this dataset, we can only compare the different estimates. The two HB estimates appear close to each other for many of the counties, with HBEP being a little closer to the direct estimates than the HBNorm. The graph clearly shows that when the sample size is small, the deviation between the direct estimates and the HB estimates is large. But as the sample size gets larger, the deviation is smaller. For the county with the largest sample size ($n_i = 802$), all the three estimates become the same.

To see the differences between the three estimators more clearly, we plotted the ratio of the DirectP and HBNorm to HBEP in the increasing order of the sample size (see Figure 3-6b). The plot confirms that when the sample size is small, the performance of the direct estimates is very variable. It also shows that except for a few counties, the two HB estimators perform more or less the same.

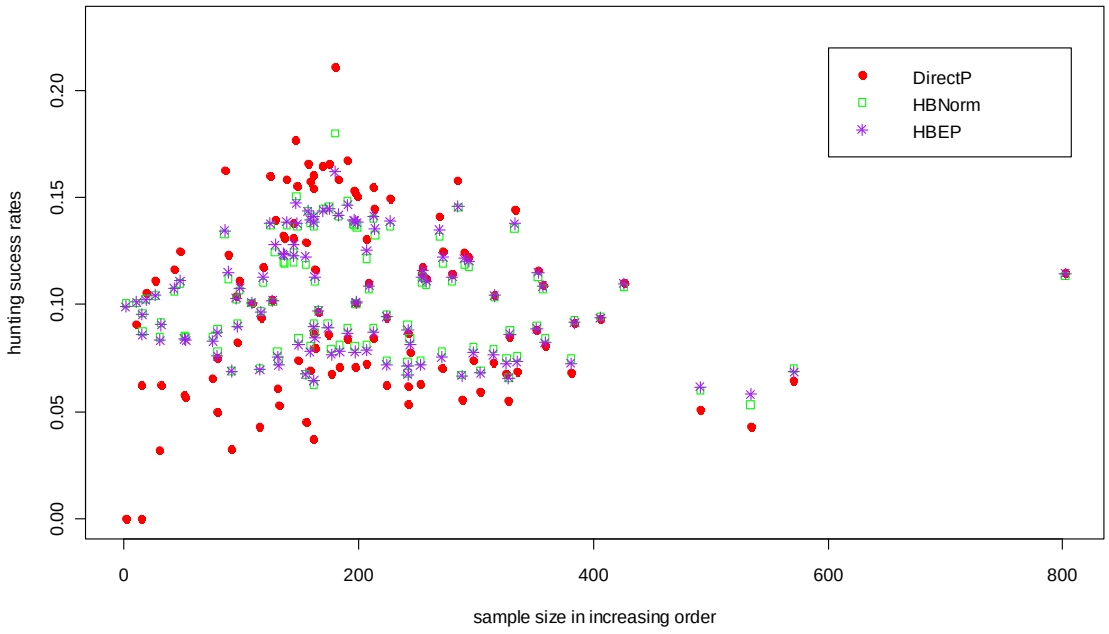

Figure 3-6a: Estimation of the Turkey hunting success rates

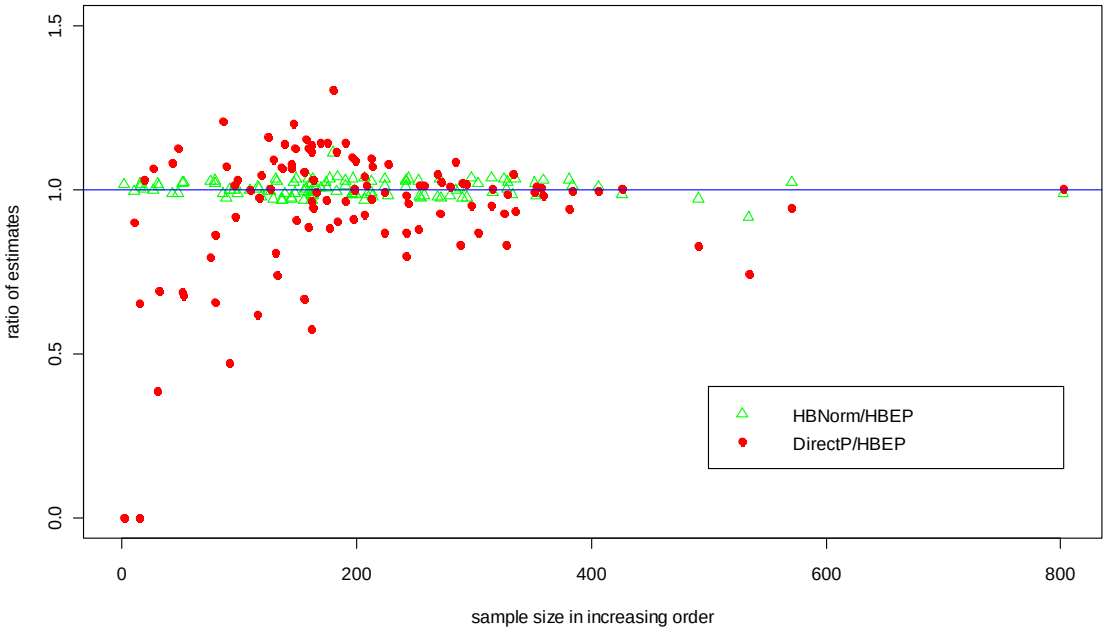

Figure 3-6b: Ratios of DirectP and HBNorm over HBEP of the hunting success rates

Figure 3-7 exhibits the standard errors/posterior standard errors associated with the estimates displayed in Figure 3-6a. The standard errors are decreasing as the sample sizes are increasing. The direct method produces extremely high standard errors for small counties. It also produces zero standard errors when the point estimates are zeros. The standard errors of the direct estimates are consistently larger than the standard errors of the two HB estimates. The standard errors of the three different estimates are getting closer as the sample sizes are increasing. We can see some differences between the posterior standard errors of the two HB estimates, but no special patterns are observed.

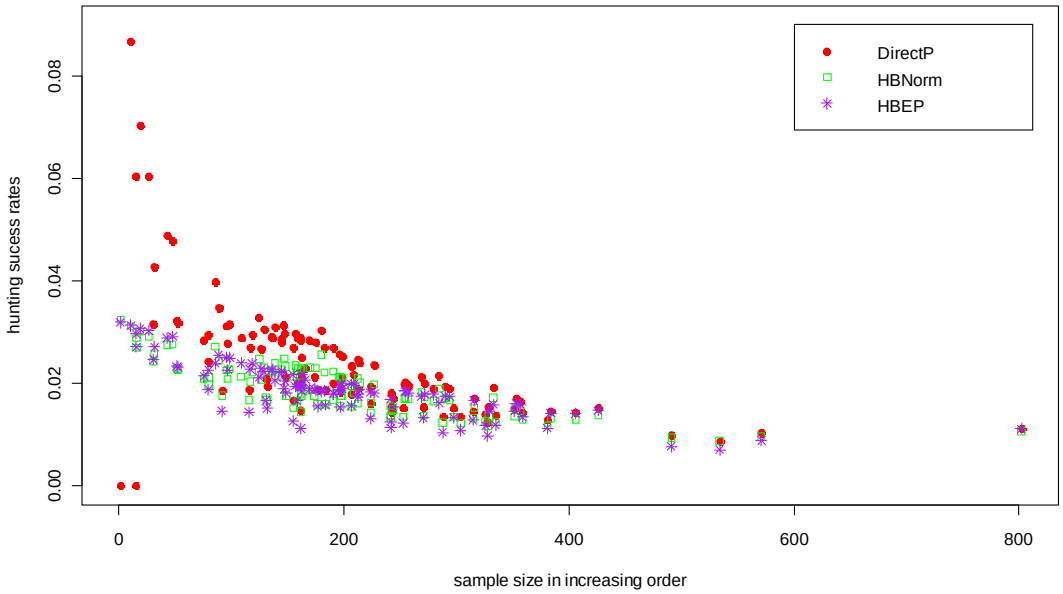

Figure 3-7: Standard errors of the direct estimates and posterior standard errors of the HB estimates of the Turkey hunting success rates

Figure 3-8 displays the posterior density plots for the hyperparameters σ_v and φ_v. The upper two panels present the standard deviation σ and the kurtosis φ_v from the *Bernoulli-Logit-EP* model respectively. The lower panel presents σ_v from the *Bernoulli-Logit-Normal* model. Both of the plots on the left show bell shapes for σ_v, although the estimates from the EP model appear to have a more sharp shape than the other one. The posterior density plot for φ_v shows that the mode of φ_v is around 0.05. The posterior mean of the kurtosis parameter φ_v is around 0.2. This evidence indicates that the random effects v_i are platykurtic and therefore the *Bernoulli-Logit-EP* model may be a more appropriate model to fit this data.

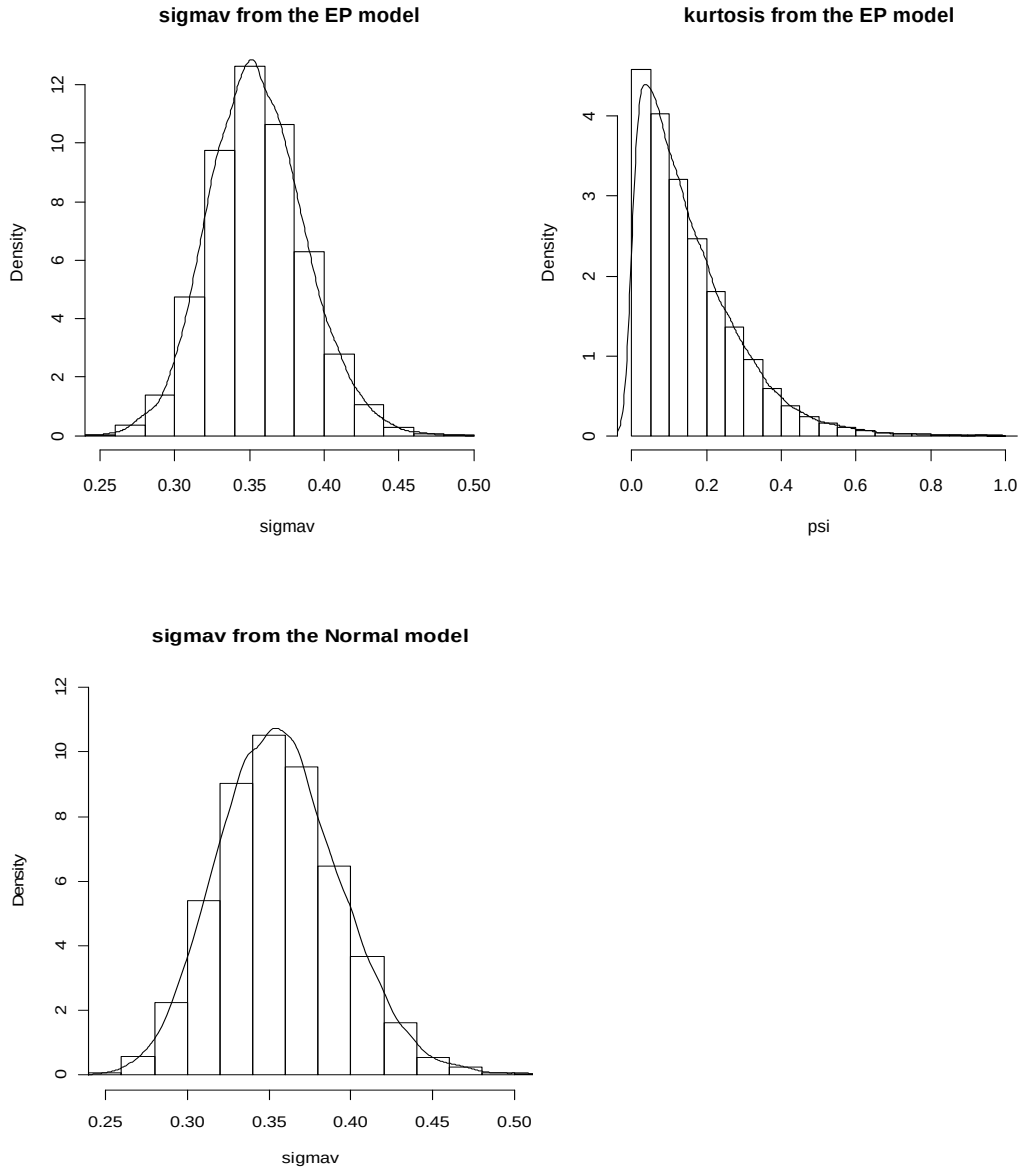

Figure 3-8: Posterior density plots of the hyperparameters σ_v and φ_v

3.7 Concluding Remarks

The proposed *Bernoulli-Logit-EP* model extends the usual logistic regression mixed model by assuming a class of probability distributions in modeling the

distribution of random effects. We considered an adaptive approach in which the shape parameter (φ_v) is determined by the survey data. The parameter φ_v is 0.5 under normality.

Our empirical data analyses based on both simulated data and real survey data demonstrate the robustness of the *Bernoulli-Logit-EP* model and suggested that the proposed model works efficiently to accommodate potential kurtosis and zero problems. To avoid computation burden, we only generated a few samples in our evaluation study based on simulated data. So the evaluation results are limited.

In this chapter, we proposed the new model for a simple sampling design from a finite population. The proposed model can be extended to accommodate multi-stage sampling designs.

Appendix for Chapter 3

Appendix A: WinBUGS code for the two HB models

A1. Code for the *Bernoulli-Logit-EP* model:

```
model
{
for ( i in 1:m)
{
yobs[i]~ dbin(theta[i], SAMPn[i])
logit(theta[i])<-inprod(beta[], X[i, ])+v[i]
}

# trick for specifying EP priors for v[i]
C<-10000
for (i in 1:m)
{
zero[i]<-0
v[i]~dunif(-10000,10000)
phi[i]<- -(log(c1)-log(sigmav)
         -pow(abs(sqrt(c0)*v[i]/sigmav),1/psi)+C
zero[i]~dpois(phi[i])
}
c0<-exp(loggam(3*psi))/exp(loggam(psi))
c1<-sqrt(c0)/(2*psi*exp(loggam(psi)))
# end of trick
for ( i in 1:p)
{
beta[i]~dflat()
}
psi~dunif(0,1)
sigmav~dunif(0, 100)
sig2v<-pow(sigmav, 2)
}
```

A2. Code for the *Bernoulli-Logit-Normal* model:

```
model
{
for ( i in 1:m)
{
yobs[i]~ dbin(theta[i], SAMPn[i])
logit(theta[i])<-inprod(beta[], X[i, ])+v[i]
```

```
v[i]~dnorm(0, precisonv)
}
for ( i in 1:p)
{
beta[i]~dflat()
}
precisonv<-1/sig2v
sig2v<-pow(sigmav, 2)
sigmav~dunif(0, 100)
}
```

Chapter 4: Bayesian Inference in Hierarchical Bayesian Models Using Approximate Methods

4.1 Introduction

We have implemented the Bayesian inference for our proposed hierarchical models using a fully Bayesian method by means of MCMC. However, some researchers have suggested alternative approximate methods such as Laplace's method and Gauss-Hermite Quadrature for Bayesian inferences. Applications of these methods were reviewed in Section 1.5. A well-known example of the use of Laplace's method is reported by Kass and Steffey (1989), who considered both first- and second-order Laplace approximations to estimate the posterior mean and posterior variance of a parameter of interest based on general conditionally independent hierarchical models. Even though Laplace's approximation can offer simple interpretations of the Bayesian methodology, the method suffers the deficiency that it is not very accurate when the sample size of the data is small. The Monte Carlo integration method is another alternative for Bayesian inference. Further, Gauss-Hermite Quadrature is also often used for numerical integration in statistics because of its relation to Gaussian densities (Liu and Pierce, 1994). To see how these methods perform for small area estimation, we study them using the proposed hierarchical Bayesian model from Chapter 3. First, we need to develop the formulas involved.

This chapter is organized as follows: The proposed hierarchical Bayesian model is represented in Section 4.2. We review different integration methods in

Section 4.3. In Section 4.4, we make Bayesian inference and conduct data analysis using simulated data assuming that all the hyperparameters are known. Section 4.5 presents the Bayesian inference when all the hyperparameters are unknown. The chapter ends with some concluding remarks in Section 4.6.

4.2 Small Area Model

Assume that a sample is drawn from a finite population using a stratified simple random sampling design. Let y_{ik} denote the binary response for a characteristic of interest for unit k in area i, where $k = 1,...,n_i$, $i = 1,...,m$. In order to accommodate kurtosis and zero problems, we apply the following robust unit level model — the *Bernoulli-Logit-EP* model defined in Chapter 3 — to estimate the area level finite population proportions $P_i = \sum_{k=1}^{N_i} y_{ik} / N_i$, $i = 1,...,m$:

$$\text{Level 1: } y_{ik} \mid \theta_i \overset{ind}{\sim} \text{Bernoulli}(\theta_i); \quad k = 1,...,n_i; \ i = 1,...,m, \tag{4.1}$$

$$\text{Level 2: } \text{logit}(\theta_i) = \mathbf{x}_i'\boldsymbol{\beta} + v_i \text{ , where } v_i \overset{iid}{\sim} EP(0, \sigma_v, \varphi_v), \ i = 1,...,m. \tag{4.2}$$

The density function of v_i is defined as:

$$f_{EP}(v_i) = \frac{c_1}{\sigma_v} \exp\left\{ -\left| \frac{\sqrt{c_0} v_i}{\sigma_v} \right|^{1/\varphi_v} \right\}, \tag{4.3}$$

where $\sigma_v \in R^+$, $\varphi_v \in (0,1]$, $c_0 = \dfrac{\Gamma(3\varphi_v)}{\Gamma(\varphi_v)}$, $c_1 = \dfrac{\sqrt{c_0}}{2\varphi_v \Gamma(\varphi_v)}$.

For the special case $\varphi_v = 0.5$, i.e., when the random effects v_i are i.i.d. normal, the model reduces to the mixed logistic regression model that has been studied in the literature (e.g., see Jiang and Lahiri, 2001).

The parameters of interest are the finite small area proportions P_i, $i = 1,...,m$. As we demonstrated in Chapter 3, inference about P_i is equivalent to inference about θ_i if $f_i \approx 0$. We will focus on the inferences about θ_i, $i = 1,...,m$. Bayesian inference for θ_i's is based on the following posterior distribution:

$$f(\theta_1,...,\theta_m \mid \mathbf{y}_S) = \int_{\boldsymbol{\beta}} \int_{\sigma_v} \int_{\varphi_v} f(\theta_1,...,\theta_m, \boldsymbol{\beta}, \sigma_v, \varphi_v \mid \mathbf{y}_S) d\boldsymbol{\beta} d\sigma_v d\varphi_v,$$

which cannot be expressed in a simple closed form. We will explore different approximate methods to estimate the posterior mean and posterior variances of the θ_i's.

4.3 Review of Various Numerical Integration Methods

Bayesian inference based on the posterior distribution of θ_i involves complex multi-dimensional integrations. Different techniques have been developed to approximate integrals in the literature. Among those, we review the first- and second-order Laplace approximations, Gauss-Hermite Quadrature, and Monte Carlo integration in this section.

First-order Laplace approximation - Assume $h(\bullet)$ is a smooth function of a $d-$dimensional parameter x with $-h(\bullet)$ having a maximum at $\hat{x}$. The Laplace method approximates an integral of the form

$$I = \int b(x) \exp\left[-nh(x)\right] dx \tag{4.4}$$

by expanding functions $h(\bullet)$ and $b(\bullet)$ around $\hat{x}$. The factor $\exp\left[-nh(x)\right]$ in the integrand is approximated by a function proportional to a normal density determined by the second-order Taylor series approximation to function $h(\bullet)$. When integrated against this normal density, the term of order $O(n^{-1/2})$ in the expansions of $b(\bullet)$ and $h(\bullet)$, which are odd functions of $x-\hat{x}$, vanish and the integral satisfies

$$I = b(\hat{x})(2\pi / n)^{d/2} \det\left[D^2 h(\hat{x})\right]^{-1/2} \exp\left[-nh(\hat{x})\right] + O(n^{-1}), \tag{4.5}$$

where $D^2 h(\hat{x})$ is the Hessian matrix of $h(\bullet)$ at $\hat{x}$ and n is the sample size. $D^a h(x)$ denotes the a-th derivative of function $h(x)$, $a \geq 1$.

Let $\mathbf{y}_s$ denote the observed data. If the posterior density of x, $f(x \mid \mathbf{y}_s)$, is proportional to $\exp[-nh(x)]$, then the posterior expectation of $b(x)$,

$$E[b(x) \mid \mathbf{y}_s] = \frac{\int b(x) f(x \mid \mathbf{y}_s) dx}{\int f(x \mid \mathbf{y}_s) dx} = \frac{\int b(x) \exp[-nh(x)] dx}{\int \exp[-nh(x)] dx} \tag{4.6}$$

may be approximated by applying the first-order Laplace method (4.5) to both the denominator and numerator to yield the *first-order* expansion

$$E[b(x) \mid \mathbf{y}_s] = b(\hat{x})[1 + O(n^{-1})]. \tag{4.7}$$

To derive the posterior variance of $b(x)$, we need the second-order Laplace approximation to $E[b(x) \mid \mathbf{y}_s]$ and $E[b^2(x) \mid \mathbf{y}_s]$, which will be introduced in the next

subsection. Kass, Tierney and Kadane (1988) obtained the *first-order* approximation to the posterior variance of $b(x)$ as follows:

$$V[b(x)\,|\,\mathbf{y}_s] = [Db(\hat{x})]'[nD^2h(\hat{x})]^{-1}[Db(\hat{x})][1+O(n^{-1})]. \tag{4.8}$$

Second-order Laplace approximation - For a one-dimensional parameter x, a second-order Laplace approximation to the integration given in (4.4) is:

$$\begin{aligned} I &= \int b(x)\exp[-nh(x)]dx \\ &= \sqrt{2\pi}\left[D^2h(\hat{x})\right]^{-1/2}\exp[-nh(\hat{x})]\Big\{b(\hat{x}) + \frac{1}{2n}\Big\{\left[D^2h(\hat{x})\right]^{-1}[D^2b(\hat{x})] - \\ &\quad \left[D^2h(\hat{x})\right]^{-2}[Db(\hat{x})][D^3h(\hat{x})] + \frac{5}{12}b(\hat{x})\left[D^3h(\hat{x})\right]^2\left[D^2h(\hat{x})\right]^{-3} \\ &\quad -\frac{1}{4}b(\hat{x})\left[D^4h(\hat{x})\right]\left[D^2h(\hat{x})\right]^2\Big\}\Big\} + O(n^{-2}). \end{aligned} \tag{4.9}$$

The posterior expectation of $b(x)$ given in (4.6) may be approximated by (4.9) to yield the *second-order* expansion:

$$\begin{aligned} E\left[b(x)\,|\,\mathbf{y}_s\right] = b(\hat{x}) &+ \frac{\left[D^2h(\hat{x})\right]^{-1}[D^2b(\hat{x})]}{2n} \\ &- \frac{\left[D^2h(\hat{x})\right]^{-2}[D^3h(\hat{x})][Db(\hat{x})]}{2n} + O(n^{-2}). \end{aligned} \tag{4.10}$$

The approximation given by (4.10) is called the *standard form* in the literature. Result (4.10) is used readily to approximate the expectation for a general function $b(x)$.

To obtain a second-order Laplace approximation to the posterior variance using the standard form, fourth and fifth derivatives of the log-likelihood of $b(x)$ would be required (Kass, Tierney and Kadane, 1988). We do not consider that approach here.

For a positive function, Tierney and Kadane (1986) obtained a *second-order* approximation to the posterior mean and variance using an alternative approach, which is described as follows. Assume $b(x)$ is a positive function. Rewrite the posterior mean of $b(x)$, $E[b(x)\,|\,\mathbf{y}_s]$, as follows:

$$E\left[b(x)\,|\,\mathbf{y}_s\right]=\frac{\int \exp\left[-n\hat{L}^*(x)\right]dx}{\int \exp\left[-n\hat{L}(x)\right]dx}, \tag{4.11}$$

where $\hat{L}(x)=-\frac{1}{n}\log\left[f(x\,|\,\mathbf{y}_s)\right]$ and $\hat{L}^*(x)=-\frac{1}{n}\log\left[b(x)\right]+\hat{L}(x)$. Assume that $\hat{x}$ and $\hat{x}^*$ maximize $-\hat{L}(x)$ and $-\hat{L}^*(x)$, respectively. Let $\tilde{\boldsymbol{\Sigma}}=\left[D^2\hat{L}(x)\right]^{-1}$ and $\tilde{\boldsymbol{\Sigma}}^*=\left[D^2\hat{L}^*(\hat{x}^*)\right]^{-1}$. Tierney and Kadane (1986) obtained the following second-order approximation to $E[b(x)\,|\,\mathbf{y}_s]$ using result (4.5):

$$E[b(x)\,|\,\mathbf{y}_s]=\left[\frac{\det\left(\tilde{\boldsymbol{\Sigma}}^*\right)}{\det\left(\tilde{\boldsymbol{\Sigma}}\right)}\right]^{1/2}\exp\left\{-n\left[\hat{L}^*(\hat{x}^*)-\hat{L}(\hat{x})\right]\right\}[1+O(n^{-2})]. \tag{4.12}$$

The form of approximation given by (4.12) is called the *fully exponential form*.

The *second-order* approximation to the posterior variance of $b(x)$ can be obtained as follows:

$$V[b(x)\,|\,\mathbf{y}_s]=\left\{E[b^2(x)\,|\,\mathbf{y}_s]-E^2[b(x)\,|\,\mathbf{y}_s]\right\}[1+O(n^{-2})], \tag{4.13}$$

where both expectation terms in (4.13) can be obtained by applying (4.12).

For more details, we refer to Kass, Tierney and Kadane (1988), and Kass and Steffey (1989).

Gauss-Hermite Quadrature (GHQ) - The GHQ formula is used to approximate an integral of the form $\int_{-\infty}^{+\infty} b(x)e^{-x^2}dx$ by the formula:

$$\int_{-\infty}^{+\infty} b(x)e^{-x^2}dx = \sum_{t=1}^{T} w_t b(x_t) + R_t, \tag{4.14}$$

where x_t are zeros of the associated Hermite polynomial $H_T(x) = (-1)^T e^{x^2}[D^T(e^{-x^2})]$ and w_t are weights defined by $w_t = \dfrac{2^{T-1}(T!)\sqrt{\pi}}{T^2[H_{T-1}(x_t)]^2}$.

Moreover, the remainder function has the form $R_T = \dfrac{T!\sqrt{\pi}}{2^T(2T!)}[D^{2T}b(\xi)]$ for some ξ, so that if $b(x)$ is a polynomial of degree $2T-1$, the remainder will be zero and the approximation becomes exact. Note that the values of x_t and w_t do not depend on the function $b(\bullet)$. Their values for specified value of T are given by Stroud and Secrest (1966, Table 5, p. 218-251).

Consider the family of Hermite formulas:

$$H_T(b) = \sum_{t=1}^{T} w_{Tt} b(x_{Tt}) \approx \int_{-\infty}^{\infty} b(x)e^{-x^2}dx. \tag{4.15}$$

If for all sufficiently large values of $|x|$, $b(x)$ satisfies the inequality $|b(x)| \le \dfrac{e^{x^2}}{|x|^{1+\rho}}$, for some $\rho > 0$, then $\lim_{T\to\infty} H_T(b) = \int_{-\infty}^{\infty} b(x)e^{-x^2}dx$. See Davis and Rabinowitz (1967, p. 96-98).

Monte Carlo integration method - Monte Carlo integration methods are algorithms for approximating any definite integrals, usually multidimensional ones. The usual algorithms evaluate the integrand on a regular grid. Monte Carlo methods randomly choose the points at which the integrand is evaluated. The traditional Monte Carlo algorithm distributes the evaluation points uniformly over the integration region. Adaptive algorithms such as VEGAS (Lepage, 1980) and MISER (Press and Farrar, 1990) use importance sampling and stratified sampling techniques to reduce the Monte Carlo error.

4.4 Bayesian Inference When $\boldsymbol{\lambda} = (\boldsymbol{\beta},\ \sigma_v,\ \varphi_v)$ is Known

The goal of this chapter is to make inference about θ_i, i.e., compute the posterior mean and posterior variance of θ_i based on model (4.1)-(4.2), by applying the techniques reviewed in Section 4.3. We start with the simple case, in which the hyperparameters $\boldsymbol{\lambda} = (\boldsymbol{\beta},\ \sigma_v,\ \varphi_v)$ are assumed known.

Let $a(\theta_i) = \text{logit}(\theta_i) - \mathbf{x}_i'\boldsymbol{\beta}$. The posterior distribution of θ_i conditioning on $\mathbf{y}_s$ and $\boldsymbol{\lambda}$ can be derived as follows:

$$
\begin{aligned}
f(\theta_i \mid \mathbf{y}_s, \boldsymbol{\lambda}) &\propto \theta_i^{\sum_{k=1}^{n_i} y_{ik}} (1-\theta_i)^{\sum_{k=1}^{n_i} (1-y_{ik})} f(\theta_i \mid \boldsymbol{\lambda}) \\
&= \theta_i^{y_i} (1-\theta_i)^{n_i - y_i} \frac{c_1}{\sigma_v \theta_i (1-\theta_i)} \exp\left[-\left| \frac{\sqrt{c_0}\, a(\theta_i)}{\sigma_v} \right|^{1/\varphi_v} \right] \\
&= \frac{c_1}{\sigma_v} \theta_i^{y_i - 1} (1-\theta_i)^{n_i - y_i - 1} \exp\left[-\frac{c_0^{1/(2\varphi_v)}}{\sigma_v^{1/\varphi_v}} |a(\theta_i)|^{1/\varphi_v} \right]. \qquad (4.16)
\end{aligned}
$$

Thus, the posterior mean of θ_i, also denoted as the best predictor (BP) of θ_i, is expressed as follows:

$$\theta_i^{BP} = E(\theta_i \mid \mathbf{y}_s, \boldsymbol{\lambda}) = \frac{\int \theta_i f(\theta_i \mid \mathbf{y}_s, \boldsymbol{\lambda}) d\theta_i}{\int f(\theta_i \mid \mathbf{y}_s, \boldsymbol{\lambda}) d\theta_i}$$
$$= \frac{\int \theta_i^{y_i} (1-\theta_i)^{n_i - y_i - 1} \exp\left[-\frac{c_0^{1/(2\varphi_v)}}{\sigma_v^{1/\varphi_v}} \left| a(\theta_i) \right|^{1/\varphi_v} \right] d\theta_i}{\int \theta_i^{y_i - 1} (1-\theta_i)^{n_i - y_i - 1} \exp\left[-\frac{c_0^{1/(2\varphi_v)}}{\sigma_v^{1/\varphi_v}} \left| a(\theta_i) \right|^{1/\varphi_v} \right] d\theta_i}. \tag{4.17}$$

The posterior variance of θ_i is expressed as:

$$V(\theta_i \mid \mathbf{y}_s, \boldsymbol{\lambda}) = E(\theta_i^2 \mid \mathbf{y}_s, \boldsymbol{\lambda}) - \left[E(\theta_i \mid \mathbf{y}_s, \boldsymbol{\lambda}) \right]^2$$
$$= \frac{\int \theta_i^2 f(\theta_i \mid \mathbf{y}_s, \boldsymbol{\lambda}) d\theta_i}{\int f(\theta_i \mid \mathbf{y}_s, \boldsymbol{\lambda}) d\theta_i} - \left(\theta_i^{BP} \right)^2. \tag{4.18}$$

The first and second moments of θ_i, $E(\theta_i \mid \mathbf{y}_s, \boldsymbol{\lambda})$ and $E(\theta_i^2 \mid \mathbf{y}_s, \boldsymbol{\lambda})$, can be obtained using the same approach. Since the integrals involved in both moments cannot be expressed in explicit form, they have to be approximated. To estimate $E(\theta_i \mid \mathbf{y}_s, \boldsymbol{\lambda})$ and $V(\theta_i \mid \mathbf{y}_s, \boldsymbol{\lambda})$, we consider the four methods reviewed in Section 4.3.

4.4.1 Approximation to the Posterior Mean and Variance of θ_i

We apply four different methods to approximate the posterior mean $E(\theta_i \mid \mathbf{y}_s, \boldsymbol{\lambda})$ and the posterior variance $V(\theta_i \mid \mathbf{y}_s, \boldsymbol{\lambda})$, starting with the first-order Laplace approximation.

First-order Laplace approximation (Method 1):

Rewrite (4.17) in the form of (4.4) as:

$$E(\theta_i \mid \mathbf{y}_S, \boldsymbol{\lambda}) = \frac{\int b(\theta_i) \exp\left[-n_i h(\theta_i)\right] d\theta_i}{\int \exp\left[-n_i h(\theta_i)\right] d\theta_i}, \tag{4.19}$$

where $b(\theta_i) = \theta_i$, and

$$h(\theta_i) = -\frac{1}{n_i}\left\{(y_i - 1)\log(\theta_i) + (n_i - y_i - 1)\log(1-\theta_i) - \frac{c_0^{1/(2\varphi_v)}}{\sigma_v^{1/\varphi_v}}\left|a(\theta_i)\right|^{1/\varphi_v}\right\}. \tag{4.20}$$

Applying (4.7), we get the first-order Laplace approximation, denoted as *LP1*, to $E\left(\theta_i \mid \mathbf{y}_S, \boldsymbol{\lambda}\right)$ as follows:

$$\hat{E}^{LP1}(\theta_i \mid \mathbf{y}_S, \boldsymbol{\lambda}) = \hat{\theta}_i + O(n_i^{-1}), \tag{4.21}$$

where $\hat{\theta}_i$ is the value which maximizes the function $-h(\theta_i)$ defined by (4.20).

To find $\hat{\theta}_i$, we need to solve the equation: $Dh(\theta_i) = 0$. That is:

$$\frac{c_0^{1/(2\varphi_v)}\left|a(\theta_i)\right|^{1/\varphi_v - 1} sign[a(\theta_i)] + \varphi_v \sigma_v^{1/\varphi_v}\left[(n_i - 2)\theta_i - y_i + 1\right]}{\varphi_v \sigma_v^{1/\varphi_v} n_i \theta_i (1-\theta_i)} = 0, \tag{4.22}$$

where $sign(\cdot)$ is the sign function defined by:

$$sign(x) = \begin{cases} \dfrac{x}{|x|}, & \text{if } x \neq 0 \\ 0, & \text{if } x = 0 \end{cases}.$$

Note that the function $\left|a(\theta_i)\right|$ is not differentiable at the point θ_i when $a(\theta_i) = 0$.

Equation (4.22) is equivalent to:

$$c_0^{1/(2\varphi_v)}\left|a(\theta_i)\right|^{1/\varphi_v - 1} sign[a(\theta_i)] + \varphi_v \sigma_v^{1/\varphi_v}\left[(n_i - 2)\theta_i - y_i + 1\right] = 0, \tag{4.23}$$

which has to be solved using a numerical method. Once the solution of equation (4.23) is obtained, substitute it into (4.21), to give the first-order Laplace approximation $\hat{E}^{LP1}(\theta_i \mid \mathbf{y}_S, \boldsymbol{\lambda})$.

Applying result (4.8), we can get the first-order Laplace approximation to the posterior variance of P_i as follows:

$$\hat{V}^{LP1}(\theta_i \mid \mathbf{y}_S, \boldsymbol{\lambda}) = \frac{\left[D^2 h(\hat{\theta}_i)\right]^{-1}}{n_i} + O(n_i^{-2}), \tag{4.24}$$

where

$$D^2 h(\hat{\theta}_i) = \frac{c_0^{1/(2\varphi_v)}(1-\varphi_v)\left|a(\hat{\theta}_i)\right|^{1/\varphi_v - 2} + \varphi_v^2 \sigma_v^{1/\varphi_v}(n_i - 2)\hat{\theta}_i(1-\hat{\theta}_i)}{\varphi_v^2 \sigma_v^{1/\varphi_v} n_i \hat{\theta}_i^2 (1-\hat{\theta}_i)^2}. \tag{4.25}$$

Note that formula (4.25) gives the exact second derivative of $h(\theta_i)$ at $\hat{\theta}_i$. It can be further simplified using asymptotic theory as follows:

$$D^2 h(\hat{\theta}_i) = \frac{1}{\hat{\theta}_i(1-\hat{\theta}_i)} + O(n_i^{-1}). \tag{4.26}$$

Substitute (4.26) into (4.24), we can get the simplified version of the first-order Laplace approximation to $V(\theta_i \mid \mathbf{y}_S, \boldsymbol{\lambda})$ as below:

$$\hat{V}^{LP1.S}(\theta_i \mid \mathbf{y}_S, \boldsymbol{\lambda}) = \frac{\hat{\theta}_i(1-\hat{\theta}_i)}{n_i} + O(n_i^{-2}). \tag{4.27}$$

From (4.21) and (4.27), the estimated posterior mean and posterior variance of θ_i are $\hat{\theta}_i$ and $\frac{\hat{\theta}_i(1-\hat{\theta}_i)}{n_i}$ respectively.

Second-order Laplace approximation (Method 2):

In this subsection, we approximate the first two posterior moments for θ_i using the second-order Laplace approximation. Applying formula (4.10), we get the the second-order Laplace approximation to $E(\theta_i \mid \mathbf{y}_s, \boldsymbol{\lambda})$ using the standard form, denoted as $SLP2$, as:

$$\hat{E}^{SLP2}(\theta_i \mid \mathbf{y}_s, \boldsymbol{\lambda}) = \hat{\theta}_i - \frac{\left[D^2 h(\hat{\theta}_i)\right]^{-2}[D^3 h(\hat{\theta}_i)]}{2n_i} + O(n_i^{-2}), \tag{4.28}$$

where $\hat{\theta}_i$ is the solution to equation (4.23), $D^2 h(\hat{\theta}_i)$ is defined by (4.25), and

$$D^3 h(\hat{\theta}_i) = \frac{c_0^{1/(2\varphi_v)}(1-\varphi_v)\left|a(\hat{\theta}_i)\right|^{1/\varphi_v - 2}\left\{(1-2\varphi_v)\left|a(\hat{\theta}_i)\right|^{-1} sign[a(\hat{\theta}_i)] + 3\varphi_v(2\hat{\theta}_i - 1)\right\}}{\varphi_v^3 \sigma^{1/\varphi_v} n_i \hat{\theta}_i^3 (1-\hat{\theta}_i)^3}$$
$$+ \frac{2(n_i - 2)(2\hat{\theta}_i - 1)}{n_i \hat{\theta}_i^2 (1-\hat{\theta}_i)^2}.$$

It is difficult to interpret (4.28) because of its complexity. For the special case, when $\varphi_v = 0.5$ (normal), we can get a simpler form of $\hat{E}^{SLP2}(\theta_i \mid \mathbf{y}_s, \boldsymbol{\lambda})$.

At $\varphi_v = 0.5$, $D^2 h(\hat{\theta}_i)$ and $D^3 h(\hat{\theta}_i)$ become:

$$D^2 h(\hat{\theta}_i) = \frac{1 + (n_i - 2)\sigma_v^2 \hat{\theta}_i (1-\hat{\theta}_i)}{n_i \sigma_v^2 \hat{\theta}_i^2 (1-\hat{\theta}_i)^2}$$

and

$$D^3 h(\hat{\theta}_i) = \frac{(2\hat{\theta}_i - 1)\left[3 + 2(n_i - 2)\sigma^2 \hat{\theta}_i (1-\hat{\theta}_i)\right]}{n_i \sigma_v^2 \hat{\theta}_i^3 (1-\hat{\theta}_i)^3}. \tag{4.29}$$

Substitute them into (4.28), we get:

$$\hat{E}^{SLP2}(\theta_i \mid \mathbf{y}_s, \boldsymbol{\lambda}) = \hat{\theta}_i - \frac{\sigma_v^2 \hat{\theta}_i(1-\hat{\theta}_i)(2\hat{\theta}_i - 1)\left[1.5 + (n_i - 2)\sigma_v^2 \hat{\theta}_i(1-\hat{\theta}_i)\right]}{\left[1 + (n_i - 2)\sigma_v^2 \hat{\theta}_i(1-\hat{\theta}_i)\right]^2} + O(n_i^{-2}). \quad (4.30)$$

There are terms of order $O(n_i^{-2})$ in the first part of formula (4.30). By ignoring them, the formula can be further simplified. We rewrite $D^3 h(\hat{\theta}_i)$ given by (4.29) as:

$$D^3 h(\hat{\theta}_i) = \frac{2(2\hat{\theta}_i - 1)}{\hat{\theta}_i^2 (1-\hat{\theta}_i)^2} + O(n_i^{-1}). \quad (4.31)$$

Now, substituting the simplified versions of $D^2 h(\hat{\theta}_i)$ given by (4.26) and of $D^3 h(\hat{\theta}_i)$ given by (4.31) into (4.28), we get the simplified version of the second-order Laplace approximation (denoted as *SLP2.s*) to $E(\theta_i \mid \mathbf{y}_s, \boldsymbol{\lambda})$ using the standard form as:

$$\hat{E}^{SLP2.s}(\theta_i \mid \mathbf{y}_s, \boldsymbol{\lambda}) = \hat{\theta}_i + \frac{1 - 2\hat{\theta}_i}{n_i} + O(n_i^{-2}). \quad (4.32)$$

The second term in (4.32) tends to zero for large sample size n_i.

Next, we provide the details for estimating $E(\theta_i \mid \mathbf{y}_s, \boldsymbol{\lambda})$ and $V(\theta_i \mid \mathbf{y}_s, \boldsymbol{\lambda})$ using the second-order Laplace approximation based on the fully exponential form. Applying result (4.12), we can get the second-order Laplace approximation to $E(\theta_i \mid \mathbf{y}_s, \boldsymbol{\lambda})$ using the fully exponential form, denoted as *FLP2*, as follows:

$$E^{FLP2}(\theta_i \mid \mathbf{y}_s, \boldsymbol{\lambda}) = \left(\frac{\tilde{\Sigma}_i^*}{\tilde{\Sigma}_i}\right)^{1/2} \exp\left\{-n_i\left[\hat{L}^*(\hat{\theta}_i^*) - \hat{L}(\hat{\theta}_i)\right]\right\}\left[1 + O(n_i^{-2})\right], \quad (4.33)$$

where

$$\begin{aligned} \hat{L}(\theta_i) &= -\frac{1}{n_i}\log\left[f(\theta_i \mid \mathbf{y}_s, \boldsymbol{\lambda})\right] \\ &= -\frac{1}{n_i}\left[(y_i - 1)\log(\theta_i) + (n_i - y_i - 1)\log(1-\theta_i) - \frac{c_0^{1/(2\varphi_v)}}{\sigma^{1/\varphi_v}}\left|a(\theta_i)\right|^{1/\varphi_v}\right]; \end{aligned} \quad (4.34)$$

$$\begin{aligned}\tilde{L}^*(\theta_i) &= -\frac{1}{n_i}\log(\theta_i) - \frac{1}{n_i}\log\left[f(\theta_i \mid \mathbf{y}_s, \boldsymbol{\lambda})\right] \\ &= -\frac{1}{n_i}\left[y_i \log(\theta_i) + (n_i - y_i - 1)\log(1-\theta_i) - \frac{c_0^{1/(2\varphi_v)}}{\sigma^{1/\varphi_v}}\left|a(\theta_i)\right|^{1/\varphi_v} \right];\end{aligned} \tag{4.35}$$

the posterior mode $\hat{\theta}_i$ is the solution to equation (4.23);

$\tilde{\Sigma}_i$ is the inverse of $D^2\tilde{L}(\hat{\theta}_i)$ which is defined by (4.25) (it is a scalar here);

the posterior mode $\hat{\theta}_i^*$ is the solution to the following equation:

$$c_0^{1/(2\varphi_v)}\left|a(\theta_i)\right|^{1/\varphi_v - 1} sign[a(\theta_i)] + \varphi_v \sigma_v^{1/\varphi_v}\left[(n_i - 1)\theta_i - y_i\right] = 0 \ ; \tag{4.36}$$

the second derivative of $\tilde{L}^*(\theta_i)$ at $\hat{\theta}_i^*$ is:

$$D^2\tilde{L}^*(\hat{\theta}_i^*) = \frac{c_0^{1/(2\varphi_v)}(1-\varphi_v)\left|a(\hat{\theta}_i^*)\right|^{1/\varphi_v - 2} + \varphi_v^2 \sigma_v^{1/\varphi_v}(n_i - 1)\hat{\theta}_i^*(1-\hat{\theta}_i^*)}{\varphi_v^2 \sigma_v^{1/\varphi_v} n_i (\hat{\theta}_i^*)^2 (1-\hat{\theta}_i^*)^2}; \tag{4.37}$$

$\tilde{\Sigma}_i^*$ is the inverse of $D^2\tilde{L}^*(\hat{\theta}_i^*)$; it is a scalar here.

To approximate $V(\theta_i \mid \mathbf{y}_s, \boldsymbol{\lambda})$, we first need to approximate $E(\theta_i^2 \mid \mathbf{y}_s, \boldsymbol{\lambda})$. Applying result (4.12) again, we get the second-order Laplace approximation using the fully exponential form to $E(\theta_i^2 \mid \mathbf{y}_s, \boldsymbol{\lambda})$ as below:

$$E^{FLP2}(\theta_i^2 \mid \mathbf{y}_s, \boldsymbol{\lambda}) = \left(\frac{\tilde{\Sigma}_i^{**}}{\tilde{\Sigma}_i}\right)^{1/2} \exp\left\{-n_i\left[\tilde{L}^{**}(\hat{\theta}_i^{**}) - \tilde{L}(\hat{\theta}_i)\right]\right\}\left[1 + O(n_i^{-2})\right], \tag{4.38}$$

where,

$$\begin{aligned}\tilde{L}^{**}(\theta_i) &= -\frac{1}{n_i}\log(\theta_i^2) - \frac{1}{n_i}\log\left[f(\theta_i \mid \mathbf{y}_s, \boldsymbol{\lambda})\right] \\ &= -\frac{1}{n_i}\left[(y_i + 1)\log(\theta_i) + (n_i - y_i - 1)\log(1-\theta_i) - \frac{c_0^{1/(2\varphi_v)}}{\sigma_v^{1/\varphi_v}}\left|a(\theta_i)\right|^{1/\varphi_v} \right]\end{aligned} \tag{4.39}$$

the posterior mode $\hat{\theta}_i^{**}$ under this case is the solution to the following equation:

$$c_0^{1/(2\varphi_v)} \left|a(\theta_i)\right|^{1/\varphi_v - 1} sign[a(\theta_i)] + \varphi_v \sigma_v^{1/\varphi_v}(n_i\theta_i - y_i - 1) = 0\,; \tag{4.40}$$

$\tilde{\Sigma}_i^{**}$ is the inverse of the second derivative of $\tilde{L}^{**}(\theta_i)$ at $\hat{\theta}_i^{**}$ and is defined as:

$$(\tilde{\Sigma}_i^{**})^{-1} = D^2\tilde{L}^*(\hat{\theta}_i^{**}) = \frac{c_0^{1/(2\varphi_v)}(1-\varphi_v)\left|a(\hat{\theta}_i^{**})\right|^{1/\varphi_v - 2} + \varphi_v^2\sigma_v^{1/\varphi_v} n_i \hat{\theta}_i^{**}(1-\hat{\theta}_i^{**})}{\varphi_v^2\sigma_v^{1/\varphi_v} n_i (\hat{\theta}_i^{**})^2 (1-\hat{\theta}_i^{**})^2}. \tag{4.41}$$

Thus, we can calculate the second order approximation of $V(\theta_i \mid \mathbf{y}_S, \boldsymbol{\lambda})$ as follows:

$$V^{FLP2}(\theta_i \mid \mathbf{y}_S, \boldsymbol{\lambda}) = \left\{\hat{E}^{FLP2}(\theta_i^2 \mid \mathbf{y}_S, \boldsymbol{\lambda}) - \left[\hat{E}^{FLP2}(\theta_i \mid \mathbf{y}_S, \boldsymbol{\lambda})\right]^2\right\}\left[1 + O(n_i^{-2})\right]. \tag{4.42}$$

Now, we want to further simplify $E^{FLP2}(\theta_i \mid \mathbf{y}_S, \boldsymbol{\lambda})$ expressed by (4.33). We showed earlier in (4.26) that $D^2\hat{L}(\hat{\theta}_i)$ can be simplified to:

$$D^2\hat{L}(\hat{\theta}_i) = \frac{1}{\hat{\theta}_i(1-\hat{\theta}_i)} + O(n_i^{-1})\,.$$

Similarly, we can simplify $D^2\tilde{L}^*(\hat{\theta}_i)$ to:

$$D^2\hat{L}(\hat{\theta}_i^*) = \frac{1}{\hat{\theta}_i^*(1-\hat{\theta}_i^*)} + O(n_i^{-1})\,.$$

Substitute the simplified versions of $D^2\tilde{L}^*(\hat{\theta}_i)$ and $D^2\hat{L}(\hat{\theta}_i)$ into (4.12), we get:

$$E^{FLP2}(\theta_i \mid \mathbf{y}_S, \boldsymbol{\lambda}) = \frac{\sqrt{\frac{1}{\hat{\theta}_i(1-\hat{\theta}_i)} + O(n_i^{-1})}\exp\left[-n_i\tilde{L}^*(\hat{\theta}_i^*)\right]}{\sqrt{\frac{1}{\hat{\theta}_i^*(1-\hat{\theta}_i^*)} + O(n_i^{-1})}\exp\left[-n_i\hat{L}(\hat{\theta}_i)\right]}\left[1 + O(n_i^{-2})\right]. \tag{4.43}$$

Substituting $\hat{L}^{*}(\hat{\theta}_i^{*})$ defined by (4.35) and $\hat{L}(\hat{\theta}_i)$ defined by (4.34) into (4.43), we get the simplified second-order approximation to $E(\theta_i \mid \mathbf{y}_s, \boldsymbol{\lambda})$ using the fully exponential form, denoted as *FLP2.s*, as:

$$\hat{E}^{FLP2.s}(\theta_i \mid \mathbf{y}_s, \boldsymbol{\lambda}) = \hat{\theta}_i^{*}\left(\frac{\hat{\theta}_i^{*}}{\hat{\theta}_i}\right)^{y_i-\frac{1}{2}}\left(\frac{1-\hat{\theta}_i^{*}}{1-\hat{\theta}_i}\right)^{n_i-y_i-\frac{1}{2}} \times \exp\left\{\frac{c_0^{1/(2\varphi_v)}}{\sigma_v^{1/\varphi_v}}\left[\left|a(\hat{\theta}_i)\right|^{1/\varphi_v} - \left|a(\hat{\theta}_i^{*})\right|^{1/\varphi_v}\right]\right\}\left[1+O(n_i^{-2})\right]. \tag{4.44}$$

Similarly, we get the simplified second-order approximation to $E(\theta_i^2 \mid \mathbf{y}_s, \boldsymbol{\lambda})$ using the fully exponential form as:

$$\hat{E}^{FLP2.s}(\theta_i^2 \mid \mathbf{y}_s, \boldsymbol{\lambda}) = \left(\hat{\theta}_i^{**}\right)^2\left(\frac{\hat{\theta}_i^{**}}{\hat{\theta}_i}\right)^{y_i-\frac{1}{2}}\left(\frac{1-\hat{\theta}_i^{**}}{1-\hat{\theta}_i}\right)^{n_i-y_i-\frac{1}{2}} \times \exp\left\{-\frac{c_0^{1/(2\varphi_v)}}{\sigma_v^{1/\varphi_v}}\left[\left|a(\hat{\theta}_i^{**})\right|^{1/\varphi_v} - \left|a(\hat{\theta}_i)\right|^{1/\varphi_v}\right]\right\}\left[1+O(n_i^{-2})\right]. \tag{4.45}$$

Substituting (4.44) and (4.45) into (4.42), we can get a simplified version of the second-order Laplace approximation to $V(\theta_i \mid \mathbf{y}_s, \boldsymbol{\lambda})$:

$$V^{FLP2.s}(\theta_i \mid \mathbf{y}_s, \boldsymbol{\lambda}) = \left\{\hat{E}^{FLP2.s}(\theta_i^2 \mid \mathbf{y}_s, \boldsymbol{\lambda}) - \left[\hat{E}^{FLP2.s}(\theta_i \mid \mathbf{y}_s, \boldsymbol{\lambda})\right]^2\right\}\left[1+O(n_i^{-2})\right]. \tag{4.46}$$

Monte Carlo method (Method 3):

From (4.2), we have $\theta_i = \dfrac{\exp(\mathbf{x}_i'\boldsymbol{\beta}+v_i)}{1+\exp(\mathbf{x}_i'\boldsymbol{\beta}+v_i)} = t(\boldsymbol{\beta}, v_i)$. Replacing θ_i in (4.17) by $t(\boldsymbol{\beta}, v_i)$, we obtain the following expression for $E(\theta_i \mid \mathbf{y}_s, \boldsymbol{\lambda})$ in terms of $\boldsymbol{\beta}$, σ_v and φ_v:

$$E(\theta_i \mid \mathbf{y}_s, \boldsymbol{\lambda}) = \frac{\int t(\boldsymbol{\beta},\ v_i)\Omega(v_i)\exp\left[-\dfrac{c_0^{1/(2\varphi_v)}}{\sigma_v^{1/\varphi_v}}|v_i|^{1/\varphi_v}\right]dv_i}{\int \Omega(v_i)\exp\left[-\dfrac{c_0^{1/(2\varphi_v)}}{\sigma_v^{1/\varphi_v}}|v_i|^{1/\varphi_v}\right]dv_i}, \tag{4.47}$$

where $\Omega(v_i) = \exp\left\{y_i v_i - n_i \log\left[1+\exp(\mathbf{x}_i'\boldsymbol{\beta}+v_i)\right]\right\}$.

The range of v_i is $(-\infty,\ +\infty)$. Let $v_i = \sigma_v \xi$, where $\xi \sim EP(0,\ 1,\ \varphi_v)$. Thus:

$$E(\theta_i \mid \mathbf{y}_s, \boldsymbol{\lambda}) = \frac{\int t(\boldsymbol{\beta},\ \sigma_v\xi)\Omega(\sigma_v\xi)\exp\left[-c_0^{1/(2\varphi_v)}|\xi|^{1/\varphi_v}\right]d\xi}{\int \Omega(\sigma_v\xi)\exp\left[-c_0^{1/(2\varphi_v)}|\xi|^{1/\varphi_v}\right]d\xi}. \tag{4.48}$$

Applying Monte Carlo integration method to (4.48), we get the Monte Carlo approximation, denoted by MC, to $E(\theta_i \mid \mathbf{y}_s, \boldsymbol{\lambda})$ as:

$$\hat{E}^{MC}(\theta_i \mid \mathbf{y}_s, \boldsymbol{\lambda}) = \frac{\sum_{r=1}^{R} t\left(\boldsymbol{\beta},\ \sigma_v\xi^{(r)}\right)\Omega\left(\sigma_v\xi^{(r)}\right)\exp\left[-c_0^{1/(2\varphi_v)}|\xi|^{1/\varphi_v}\right]}{\sum_{r=1}^{R}\Omega\left(\sigma_v\xi^{(r)}\right)\exp\left[-c_0^{1/(2\varphi_v)}|\xi|^{1/\varphi_v}\right]}, \tag{4.49}$$

where $\xi^{(r)} \sim EP(0,\ 1,\ \varphi_v),\quad \mathrm{r} = 1,\ldots,R.$

Similarly, we can estimate $E(\theta_i^2 \mid \mathbf{y}_s, \boldsymbol{\lambda})$ as:

$$\hat{E}^{MC}(\theta_i^2 \mid \mathbf{y}_s, \boldsymbol{\lambda}) = \frac{\sum_{r=1}^{R} t^2\left(\boldsymbol{\beta},\ \sigma_v\xi^{(r)}\right)\Omega\left(\sigma_v\xi^{(r)}\right)\exp\left[-c_0^{1/(2\varphi_v)}\left|\xi^{(r)}\right|^{1/\varphi_v}\right]}{\sum_{r=1}^{R}\Omega\left(\sigma_v\xi^{(r)}\right)\exp\left[-c_0^{1/(2\varphi_v)}\left|\xi^{(r)}\right|^{1/\varphi_v}\right]}. \tag{4.50}$$

Therefore, we can obtain the Monte Carlo approximation to $V(\theta_i \mid \mathbf{y}_s, \boldsymbol{\lambda})$ as:

$$\hat{V}^{MC}(\theta_i \mid \mathbf{y}_s, \boldsymbol{\lambda}) = \hat{E}^{MC}(\theta_i^2 \mid \mathbf{y}_s, \boldsymbol{\lambda}) - \left[\hat{E}^{MC}(\theta_i \mid \mathbf{y}_s, \boldsymbol{\lambda})\right]^2, \tag{4.51}$$

where the first and second terms in (4.51) are defined by (4.50) and (4.49) respectively.

Numerical Integration using GHQ (Method 4):

Let $\xi = \sqrt{2}\zeta$. We rewrite (4.48) in terms of ζ as follows:

$$E(\theta_i \mid \mathbf{y}_s, \boldsymbol{\lambda}) = \frac{\int t\left(\boldsymbol{\beta},\ \sqrt{2}\sigma_v\zeta\right)\Omega\left(\sqrt{2}\sigma_v\zeta\right)\exp\left[-(2c_0)^{1/2\varphi_v}\left|\zeta\right|^{1/\varphi_v} + \zeta^2\right]\exp\left(-\zeta^2\right)d\zeta}{\int \Omega\left(\sqrt{2}\sigma_v\zeta\right)\exp\left[-(2c_0)^{1/2\varphi_v}\left|\zeta\right|^{1/\varphi_v} + \zeta^2\right]\exp\left(-\zeta^2\right)d\zeta}. \tag{4.52}$$

Applying GHQ to (4.52), we obtain the numerical GHQ approximation to $E(\theta_i \mid \mathbf{y}_s, \boldsymbol{\lambda})$ as follows:

$$\hat{E}^{GHQ}(\theta_i \mid \mathbf{y}_s, \boldsymbol{\lambda}) = \frac{\sum_{k=1}^{K} w_k t(\boldsymbol{\beta},\ \sqrt{2}\sigma_v\zeta_k)\Omega(\sqrt{2}\sigma_v\zeta_k)\exp\left[-(2c_0)^{1/2\varphi}\left|\zeta_k\right|^{\frac{1}{\varphi}} + \zeta_k^2\right]}{\sum_{k=1}^{K} w_k \Omega(\sqrt{2}\sigma_v\zeta_k)\exp\left[-(2c_0)^{1/2\varphi}\left|\zeta_k\right|^{\frac{1}{\varphi}} + \zeta_k^2\right]}, \tag{4.53}$$

where K is the number of quadrature points, w_k is the quadrature weight, and ζ_k is the quadrature node.

Note that function *gqz*( ) in R gives the quadrature weight and node for a given number of quadrature points for the integral of the form $\int_{-\infty}^{+\infty} f(t)e^{-t^2/2}dt$. To use the *gqz*( ) function, we do not need the $\xi = \sqrt{2}\zeta$ transformation.

Similarly, we can get the numerical GHQ approximation to $E(\theta_i^2 \mid \mathbf{y}_s, \boldsymbol{\lambda})$ as:

$$\hat{E}^{GHQ}(\theta_i^2 \mid \mathbf{y}_s, \boldsymbol{\lambda}) = \frac{\sum_{k=1}^{K} w_k t^2\left(\boldsymbol{\beta},\ \sqrt{2}\sigma_v\zeta_k\right)\Omega\left(\sqrt{2}\sigma_v\zeta_k\right)\exp\left[-(2c_0)^{1/(2\varphi_v)}\left|\zeta_k\right|^{1/\varphi_v} + \zeta_k^2\right]}{\sum_{k=1}^{K}\Omega\left(\sqrt{2}\sigma_v\zeta_k\right)w_k \exp\left[-(2c_0)^{1/(2\varphi_v)}\left|\zeta_k\right|^{1/\varphi_v} + \zeta_k^2\right]}.$$

The numerical GHQ approximation to $V(\theta_i \mid \mathbf{y}_s, \boldsymbol{\lambda})$ can be obtained using the following:

$$\hat{V}^{GHQ}(\theta_i \mid \mathbf{y}_s, \boldsymbol{\lambda}) = \hat{E}^{GHQ}(\theta_i^2 \mid \mathbf{y}_s, \boldsymbol{\lambda}) - \left[\hat{E}^{GHQ}(\theta_i \mid \mathbf{y}_s, \boldsymbol{\lambda})\right]^2. \qquad (4.54)$$

4.4.2 Data Analysis Using Simulated Data

In this data analysis, we consider data generated using three scenarios for the kurtosis φ: i) $\varphi_v = 0.5$ (normal); ii) $\varphi_v = 0.2$ (platykurtic); iii) $\varphi_v = 0.8$ (leptokurtic). Our objective is to estimate $E(\theta_i \mid \mathbf{y}_s, \boldsymbol{\lambda})$ and $V(\theta_i \mid \mathbf{y}_s, \boldsymbol{\lambda})$ using the methods described in Section 4.3.1.

For each value of φ_v, we generated the true θ_i and observed y_i using model (4.1)~(4.2) for $m = 18$ small areas based on the following information which was based on the baseball data analized in Chapter 3:

i) $\sigma_v = 0.1654$;

ii) $\boldsymbol{\beta} = (-1.716,\ 2.703)'$;

iii) $\mathbf{x} = (0.118,\ 0.249,\ 0.246,\ 0.264,\ 0.314,\ 0.275,\ 0.255,\ 0.248,\ 0.256,$
$0.255,\ 0.303,\ 0.234,\ 0.281,\ 0.250,\ 0.244,\ 0.244,\ 0.257,\ 0.271)$;

iv) $n_i = 45,\ i = 1,...,18$.

Using the observed data y_i, we first obtained the posterior distribution $f(\theta_i \mid \mathbf{y}_s, \boldsymbol{\lambda})$ using result (4.16). Figure 4-1 illustrates the plots of the posterior distribution $f(\theta_1 \mid \mathbf{y}_s, \boldsymbol{\lambda})$ for the three different φ_v values using the data from the first small area. Figures 4-2 presents the corresponding plots for $\log\left[f(\theta_1 \mid \mathbf{y}_s, \boldsymbol{\lambda})\right]$.

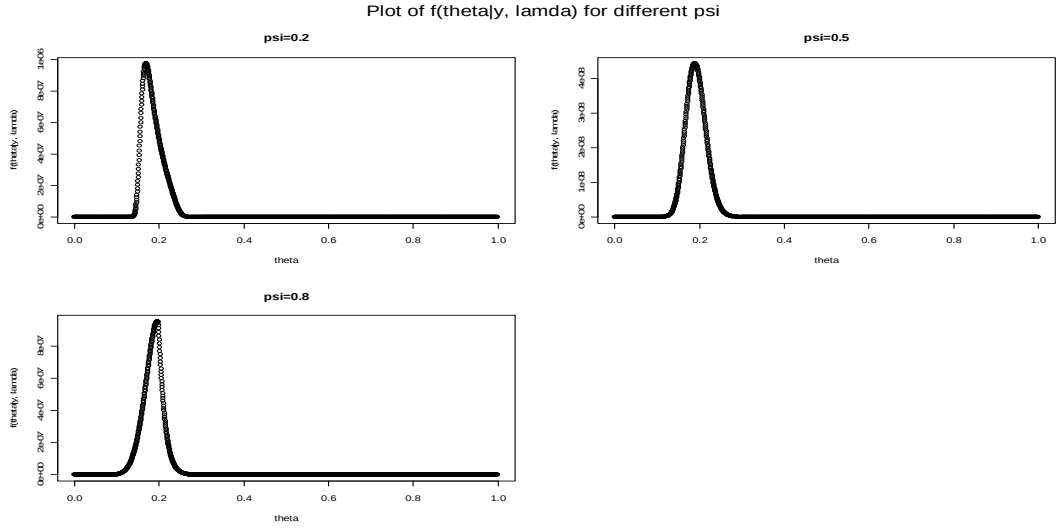

Figure 4-1: The distribution of $f(\theta_1 \mid \mathbf{y}_s, \boldsymbol{\lambda})$ for the three values of φ_v

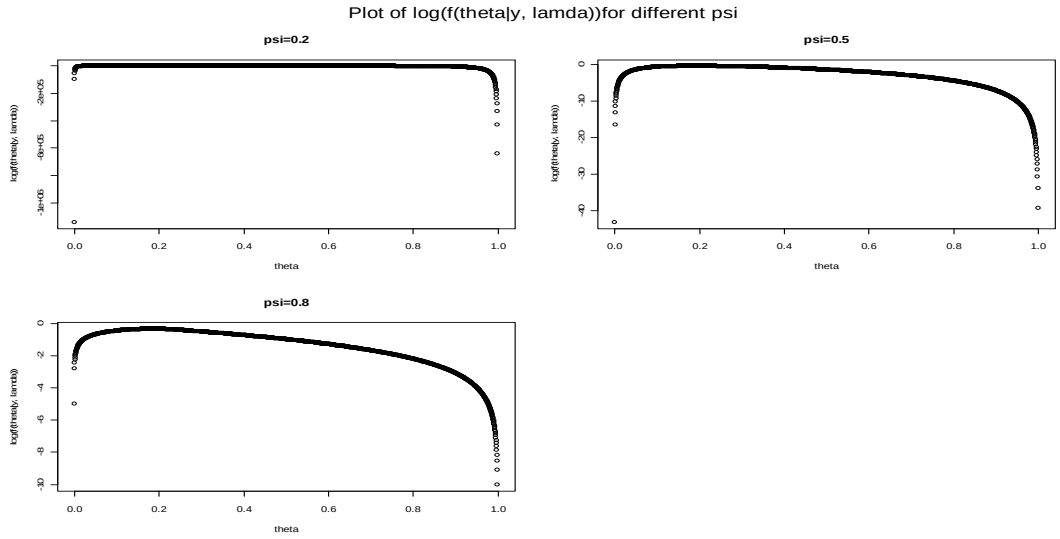

Figure 4-2: The distribution of $\log\left[f(\theta_1 \mid \mathbf{y}_s, \boldsymbol{\lambda})\right]$ for the three values of φ_v

We then approximated $E(\theta_i \mid \mathbf{y}_s, \boldsymbol{\lambda})$ using the GHQ method defined by (4.53), the Monte Carlo (MC) method defined by (4.49), the first-order Laplace approximation (LP1) defined by (4.21), the second-order Laplace approximation, which includes four versions: the standard form (SLP2) defined by (4.28); the fully exponential form (FLP2) defined by (4.33); the simplified standard form (SLP2.s) defined by (4.32), and the simplified fully exponential form (FLP2.s) defined by (4.44). We also implemented the MCMC method described in Section 1.5.2 for evaluation purposes. Among these methods, the GHQ method should give the most accurate results since it uses purely numerical approximation. To compare different methods, we treated the values for the 18 areas given by the GHQ method as the gold standard values, and computed the following summary statistics for all other methods:

- Average absolute deviation (AAD), $AAD = \frac{1}{m}\sum_{i=1}^{m}\left|est(\theta_i) - \theta_i^{GHQ}\right|$
- Average absolute relative deviation (AARD),

$$AARD = \frac{1}{m}\sum_{i=1}^{m}\left|est(\theta_i) - \theta_i^{GHQ}\right| \Big/ \theta_i^{GHQ}$$

where θ_i^{GHQ} and $est(\theta_i)$ denote the estimate of $E(\theta_i \mid \mathbf{y}_s, \boldsymbol{\lambda})$ using GHQ and other estimation methods respectively.

Tables 4-1 presents the summary statistics for the six different approximations to the posterior mean $E(\theta_i \mid \mathbf{y}_s, \boldsymbol{\lambda})$ when the random effect v_i is normal ($\varphi_v = 0.5$), platykurtic ($\varphi_v = 0.2$), and leptokurtic ($\varphi_v = 0.8$) respectively. The results show some inconsistent patterns for the Laplace's method. For example, when v_i is platykurtic or leptokurtic, the first-order Laplace (LP1) performs better than the

second-order Laplace (SLP2 and FLP2), probably because the sample sizes n_i are small. In addition, the simplified versions of the Laplace's estimates (SLP2.s and FLP2.s) did not show consistent improvement over SLP2 and FLP2. The patterns between MC and the Laplace's method are not consistent as φ_v varies. The MCMC method performed consistently better than the other methods in the non-normal cases.

Table 4-1: Summary statistics AAD ($\times 10^{-4}$) and AARD ($\times 10^{-4}$) for different approximations to the posterior mean of θ_i given different values of φ_v and $n_i = 45$

	$\varphi_v = 0.5$ (normal)		$\varphi_v = 0.2$ (platykurtic)		$\varphi_v = 0.8$ (leptokurtic)	
	AAD	AARD	AAD	AARD	AAD	AARD
MCMC	1.5	6.0	1.9	7.4	2.2	8.6
MC	42.5	161.5	40.2	157.1	79.6	299.1
LP1	28.5	109.8	134.7	523.2	100.3	374.4
SLP2	0.4	1.4	330.4	1,217.6	176.9	814.9
FLP2	0.2	0.8	317.1	1,173.7	421.1	1,636.9
SLP2.s	78.4	304.2	122.9	452.0	109.6	432.2
FLP2.s	2.5	9.9	137.3	527.8	97.9	369.4

Note: the GHQ defined by (4.53) is used as gold standard to compute the summary statistics for the methods of: Markov Chain Monte Carlo (MCMC), Monte Carlo (MC) defined by (4.49), first-order Laplace (LP1) defined by (4.21), second-order Laplace using the standard form (SLP2) defined by (4.28) and the fully exponential form (FLP2) defined by (4.33), simplified SLP2 (SLP2.s) defined by (4.32), and simplified FLP2 (FLP2.s) defined by (4.44).

We also estimated the posterior variance $V(\theta_i \mid \mathbf{y}_s, \boldsymbol{\lambda})$ using MCMC, the numerical GHQ method defined by (4.54), the Monte Carlo (MC) method defined by (4.51), the first-order Laplace approximation (LP1) method defined by (4.24) and its simplified formula (LP1.s) defined by (4.27), the second-order Laplace approximation method using the fully exponential form (FLP2) defined by (4.42) and the simplified formula (FLP2.s) defined by (4.46). Again we treated the numerical

GHQ approximation values as the standard values and computed the summary statistics AAD and AARD for all other methods. Corresponding to Table 4-1, the results for approximating the posterior variance are presented in Table 4-2. The results show similar patterns to those in Table 4-1.

Table 4-2: Summary statistics AAD ($\times 10^{-6}$) and AARD ($\times 10^{-6}$) for different approximations to the posterior variance of θ_i given different values of φ_v and $n_i = 45$

	$\varphi_v = 0.5$ (normal)		$\varphi_v = 0.2$ (platykurtic)		$\varphi_v = 0.8$ (leptokurtic)	
	AAD	AARD	AAD	AARD	AAD	AARD
MCMC	6.4	7,925.2	9.2	10,969.8	16.7	19,881.9
MC	404.6	492,797.1	311.5	366,589.1	563.1	623,880.4
LP1	2.5	3,151.4	1,128.7	1,216,184.0	606.8	720,160.7
FLP2	0.1	105.8	7,212.2	7,891,897.0	17,089.9	18,410,918.2
LP1.s	3,439.8	4,215,435.5	3,420.4	4,160,449.3	3,421.4	4,042,930.5
FLP2.s	4.9	5,842.6	861.7	935,505.1	570.7	700,210.7

Note: the GHQ defined by (4.54) is used as gold standard to compute the summary statistics for the methods of: Markov Chain Monte Carlo (MCMC), Monte Carlo (MC) defined by (4.51), first-order Laplace (LP1) defined by (4.24), second-order Laplace using the fully exponential form (FLP2) defined by (4.42), simplified LP1 (LP1.s) defined by (4.27), and simplified FLP2 (FLP2.s) defined by (4.46).

Tables 4-1 and 4-2 have demonstrated that the MCMC method performs closest to the GHQ method compared to other approximation methods in the non-normal cases. However, MCMC method performs secondary to the second-order Laplace method for the normal cases. The results for Laplace approximations are inconsistent; sometimes the first-order approximation performs better than the second-order approximation. This may due to the reminder terms $O(n_i^{-1})$ and

$O(n_i^{-2})$, which are not negligible for small n_i, even though the sample size $n_i = 45$ in this exercise is not small in a typical small area estimation problem. To confirm this, we replicate the study using data generated with large sample size $n_i = 1,000$. The summary results are presented in Tables 4-3 and 4-4.

Table 4-3: Summary statistics AAD ($\times 10^{-4}$) and AARD ($\times 10^{-4}$) for different approximations to the posterior mean of θ_i given different values of φ_v and $n_i = 1,000$

	$\varphi_v = 0.5$ (normal)		$\varphi_v = 0.2$ (platykurtic)		$\varphi_v = 0.8$ (leptokurtic)	
	AAD	AARD	AAD	SRASRD	AAD	SRASRD
MCMC	1.3	4.8	2.8	9.3	2.0	7.6
MC	40.9	158.4	42.6	156.7	45.4	173.5
LP1	4.2	16.4	9.0	36.9	8.2	31.7
SLP2	1.0	3.7	3.1	10.6	7.5	29.3
FLP2	1.0	3.7	3.1	10.8	6.4	24.8
SLP2.s	1.1	3.8	7.9	30.9	9.1	36.3
FLP2.s	1.1	4.0	8.0	31.4	8.7	35.0

Note: the GHQ defined by (4.53) is used as gold standard to compute the summary statistics for the methods of: Markov Chain Monte Carlo (MCMC), Monte Carlo (MC) defined by (4.49), first-order Laplace (LP1) defined by (4.21), second-order Laplace using the standard form (SLP2) defined by (4.28) and the fully exponential form (FLP2) defined by (4.33), simplified SLP2 (SLP2.s) defined by (4.32), and simplified FLP2 (FLP2.s) defined by (4.44).

Table 4-4: Summary statistics AAD ($\times 10^{-6}$) and AARD ($\times 10^{-6}$) for different approximations to the posterior variance of θ_i given different values of φ_v and $n_i = 1,000$

	$\varphi_v = 0.5$ (normal)		$\varphi_v = 0.2$ (platykurtic)		$\varphi_v = 0.8$ (leptokurtic)	
	AAD	AARD	AAD	AARD	AAD	AARD
MCMC	4.3	26,106.9	7.8	53,206.8	4.5	28,160.6
MC	37.4	231,991.3	31.7	203,207.4	46.3	264,973.3
LP1	4.9	29,512.2	11.9	79,032.4	21.1	138,137.6
FLP2	4.9	29,381.1	7.9	53,498.2	102.2	713,439.9
LP1.s	31.5	195,144.6	35.2	257,562.1	27.0	178,439.1
FLP2.s	4.9	29,504.5	11.9	79,647.6	25.2	163,313.0

Note: the GHQ defined by (4.54) is used as gold standard to compute the summary statistics for the methods of: Markov Chain Monte Carlo (MCMC), Monte Carlo (MC) defined by (4.51), first-order Laplace (LP1) defined by (4.24), second-order Laplace using the fully exponential form (FLP2) defined by (4.42), simplified LP1 (LP1.s) defined by (4.27), and simplified FLP2 (FLP2.s) defined by (4.46).

From Tables 4-3 and 4-4, we can see the results for Laplace's method are as expected, with the second-order approximation working consistently better than the first-order approximation. The second-order Laplace performs closer to the MCMC method with these large samples. The MCMC method still performs the closest to the GHQ method among all the approximation methods in the non-normal cases. In the normal cases, the performances of MCMC method and the second-order Laplace method are very close now. There is no evidence showing that the simplified versions of the Laplace approximations are consistently better than the non-simplified versions.

4.5 Bayesian Inference when $\boldsymbol{\lambda} = (\boldsymbol{\beta},\ \sigma_v,\ \varphi_v)$ is Unknown

In practice, the hyperparameter vector $\boldsymbol{\lambda}$ is unknown**.** A prior assumption $\pi(\boldsymbol{\lambda})$ is often applied in Bayesian analysis. This section studies how to make inference about θ_i based on model (4.1)~(4.2) when the hyperparameter vector $\boldsymbol{\lambda}$ is unknown.

4.5.1 Bayesian Inference for a General Function of θ_i

Let $b(\theta_i)$ be a continuous function of θ_i having the first three derivatives. Our goal is to estimate the posterior mean $E\left[b(\theta_i) \mid \mathbf{y}_s\right]$ and posterior variance $V\left[b(\theta_i) \mid \mathbf{y}_s\right]$**.**

Note that:

$$E\left[b(\theta_i) \mid \mathbf{y}_s\right] = E_{\boldsymbol{\lambda}}\left\{E\left[b(\theta_i) \mid \mathbf{y}_s, \boldsymbol{\lambda}\right] \middle| \mathbf{y}_s\right\}; \tag{4.55}$$

$$V\left[b(P_i) \mid \mathbf{y}_s\right] = E_{\boldsymbol{\lambda}}\left\{V\left[b(\theta_i) \mid \mathbf{y}_s, \boldsymbol{\lambda}\right] \middle| \mathbf{y}_s\right\} + V_{\boldsymbol{\lambda}}\left\{E\left[b(\theta_i) \mid \mathbf{y}_s, \boldsymbol{\lambda}\right] \middle| \mathbf{y}_s\right\}. \tag{4.56}$$

Assume that $E\left[b(\theta_i) \mid \mathbf{y}_s, \boldsymbol{\lambda}\right]$ and $V\left[b(\theta_i) \mid \mathbf{y}_s, \boldsymbol{\lambda}\right]$ can be written as functions of $\boldsymbol{\lambda}$ analytically, that is, $E\left[b(\theta_i) \mid \mathbf{y}_s, \boldsymbol{\lambda}\right] = G(\boldsymbol{\lambda})$ and $V\left[b(\theta_i) \mid \mathbf{y}_s, \boldsymbol{\lambda}\right] = H(\boldsymbol{\lambda})$, where G and H are some smooth functions of $\boldsymbol{\lambda}$ having the first three derivatives. Then (4.55) and (4.56) become:

$$E\left[b(\theta_i) \mid \mathbf{y}_s\right] = E_{\boldsymbol{\lambda}}\left[G(\boldsymbol{\lambda}) \mid \mathbf{y}_s\right]; \tag{4.57}$$

$$\begin{aligned} V\left[b(\theta_i) \mid \mathbf{y}_s\right] &= E_{\boldsymbol{\lambda}}\left[H(\boldsymbol{\lambda}) \mid \mathbf{y}_s\right] + V_{\boldsymbol{\lambda}}\left[G(\boldsymbol{\lambda}) \mid \mathbf{y}_s\right] \\ &= E_{\boldsymbol{\lambda}}\left[H(\boldsymbol{\lambda}) \mid \mathbf{y}_s\right] + E_{\boldsymbol{\lambda}}\left[G^2(\boldsymbol{\lambda}) \mid \mathbf{y}_s\right] - \left\{E_{\boldsymbol{\lambda}}\left[G(\boldsymbol{\lambda}) \mid \mathbf{y}_s\right]\right\}^2. \end{aligned} \tag{4.58}$$

Next, we illustrate how to approximate $E_{\boldsymbol{\lambda}}\left[G(\boldsymbol{\lambda})\,|\,\mathbf{y}_s\right]$. A Similar approach can be used to approximate each term in (4.58).

Since $E_{\boldsymbol{\lambda}}\left[G(\boldsymbol{\lambda})\,|\,\mathbf{y}_s\right]=\dfrac{\int G(\boldsymbol{\lambda})f(\boldsymbol{\lambda}\,|\,\mathbf{y}_s)d\boldsymbol{\lambda}}{\int f(\boldsymbol{\lambda}\,|\,\mathbf{y}_s)d\boldsymbol{\lambda}}$, we need to find the posterior distribution $f(\boldsymbol{\lambda}\,|\,\mathbf{y}_s)$. The joint density of $\mathbf{y}_s$, $\boldsymbol{\lambda}$ and $\mathbf{v}=(v_1,...,v_m)'$ is:

$$
\begin{aligned}
f(\mathbf{y}_s,\,\boldsymbol{\lambda},\,\mathbf{v}) &\propto f(\mathbf{y}_s\,|\,\mathbf{v},\,\boldsymbol{\lambda})f(\mathbf{v}\,|\,\boldsymbol{\lambda})\pi(\boldsymbol{\lambda}) \\
&\propto \prod_{i=1}^{m}\left\{\left[\frac{\exp(\mathbf{x}_i'\boldsymbol{\beta}+v_i)}{1+\exp(\mathbf{x}_i'\boldsymbol{\beta}+v_i)}\right]^{y_i}\left[\frac{1}{1+\exp(\mathbf{x}_i'\boldsymbol{\beta}+v_i)}\right]^{n_i-y_i}\frac{c_1}{\sigma_v}H(v_i,\boldsymbol{\lambda})\right\}\pi(\boldsymbol{\lambda}) \\
&= c_1^m\sigma_v^{-m}\pi(\boldsymbol{\lambda})\prod_{i=1}^{m}\left\{\left[\frac{\exp(\mathbf{x}_i'\boldsymbol{\beta}+v_i)}{1+\exp(\mathbf{x}_i'\boldsymbol{\beta}+v_i)}\right]^{y_i}\left[\frac{1}{1+\exp(\mathbf{x}_i'\boldsymbol{\beta}+v_i)}\right]^{n_i-y_i}H(v_i,\boldsymbol{\lambda})\right\},
\end{aligned}
$$

where $H(v_i,\boldsymbol{\lambda})=\exp\left[-\dfrac{c_0^{1/(2\varphi_v)}}{\sigma_v^{1/\varphi_v}}|v_i|^{1/\varphi_v}\right]$. Note that both c_0 and c_1 are functions of φ_v. Thus, the joint density of $\mathbf{y}_s$ and $\boldsymbol{\lambda}$ can be derived as below:

$$
\begin{aligned}
&f(\mathbf{y}_s,\,\boldsymbol{\lambda})=\int f(\mathbf{y}_s,\,\boldsymbol{\lambda},\,\mathbf{v})d\mathbf{v} \\
&\propto c_1^m\sigma_v^{-m}\pi(\boldsymbol{\lambda})\int\prod_{i=1}^{m}\left\{\left[\frac{\exp(\mathbf{x}_i'\boldsymbol{\beta}+v_i)}{1+\exp(\mathbf{x}_i'\boldsymbol{\beta}+v_i)}\right]^{y_i}\left[\frac{1}{1+\exp(\mathbf{x}_i'\boldsymbol{\beta}+v_i)}\right]^{n_i-y_i}H(v_i,\boldsymbol{\lambda})\right\}d\mathbf{v} \\
&= c_1^m\sigma_v^{-m}\pi(\boldsymbol{\lambda})\int...\int\prod_{i=1}^{m}\left\{\frac{\exp\left[y_i(\mathbf{x}_i'\boldsymbol{\beta}+v_i)\right]}{\left[1+\exp(\mathbf{x}_i'\boldsymbol{\beta}+v_i)\right]^{n_i}}H(v_i,\boldsymbol{\lambda})\right\}dv_1...dv_m \\
&= c_1^m\sigma_v^{-m}\pi(\boldsymbol{\lambda})\prod_{i=1}^{m}\int\frac{\exp\left[y_i(\mathbf{x}_i'\boldsymbol{\beta}+v_i)\right]}{\left[1+\exp(\mathbf{x}_i'\boldsymbol{\beta}+v_i)\right]^{n_i}}H(v_i,\boldsymbol{\lambda})dv_i \\
&= c_1^m\sigma_v^{-m}\pi(\boldsymbol{\lambda})\prod_{i=1}^{m}\int\left\{\frac{\exp\left[y_i(\mathbf{x}_i'\boldsymbol{\beta}+v_i)-\dfrac{c_0^{1/(2\varphi_v)}}{\sigma_v^{1/\varphi_v}}|v_i|^{1/\varphi_v}\right]}{\left[1+\exp(\mathbf{x}_i'\boldsymbol{\beta}+v_i)\right]^{n_i}}\right\}dv_i.
\end{aligned}
\tag{4.59}
$$

Note that at the penultimate step of (4.59), we are able to move the product over the small areas (the index i) to the outside of the integral because the random effects v_i are assumed independent. Since the integral with the product over the index i has no closed-form, numerical integration has to be used.

Applying the numerical GHQ method to the integral in (4.59), we get:

$$f(\mathbf{y}_S,\ \boldsymbol{\lambda}) = \int f(\mathbf{y}_S,\ \boldsymbol{\lambda},\ \mathbf{v})\mathrm{d}\mathbf{v}$$

$$\propto c_1^m \sigma_v^{-m} \pi(\boldsymbol{\lambda}) \prod_{i=1}^{m} \sum_{t=1}^{T} \left\{ \frac{w_t \exp\left[y_i(\mathbf{x}_i'\boldsymbol{\beta} + v_t) - \dfrac{c_0^{1/(2\varphi_v)}}{\sigma_v^{1/\varphi_v}} \left| v_t \right|^{1/\varphi_v} + v_t^2 \right]}{\left[1 + \exp(\mathbf{x}_i'\boldsymbol{\beta} + v_t)\right]^{n_i}} \right\},$$

where T is the number of quadrature points , w_t is the quadrature weight, and v_t is the quadrature node for a given number of quadrature points, $t = 1,...,T$. Thus the posterior distribution of $\boldsymbol{\lambda}$ is given by:

$$f(\boldsymbol{\lambda} \mid \mathbf{y}_S) \propto c_1^m \sigma^{-m} \pi(\boldsymbol{\lambda}) \prod_{i=1}^{m} \sum_{t=1}^{T} \left\{ \frac{w_t \exp\left[y_i(\mathbf{x}_i'\boldsymbol{\beta} + v_t) - \dfrac{c_0^{1/(2\varphi_v)}}{\sigma_v^{1/\varphi_v}} \left| v_t \right|^{1/\varphi_v} + v_t^2 \right]}{\left[1 + \exp(\mathbf{x}_i'\boldsymbol{\beta} + v_t)\right]^{n_i}} \right\}. \quad (4.60)$$

Note that (4.60) only holds if the term on the right hand side is proper, that is, the multi-dimensional integrals over $\boldsymbol{\lambda}$ for the term on the right hand side of (4.60) must be bounded.

As a result, we can write the posterior expectation of $G(\boldsymbol{\lambda})$ as:

$$E\left[G(\boldsymbol{\lambda}) \mid \mathbf{y}_S\right] = \frac{\int G(\boldsymbol{\lambda}) f(\boldsymbol{\lambda} \mid \mathbf{y}_S) d\boldsymbol{\lambda}}{\int f(\boldsymbol{\lambda} \mid \mathbf{y}_S) d\boldsymbol{\lambda}} = \frac{\int G(\boldsymbol{\lambda}) \exp\left[-mh(\boldsymbol{\lambda})\right] d\boldsymbol{\lambda}}{\int \exp\left[-mh(\boldsymbol{\lambda})\right] d\boldsymbol{\lambda}}, \quad (4.61)$$

where

$$h(\boldsymbol{\lambda}) = -\frac{1}{m}\log\left[f(\boldsymbol{\lambda}\,|\,\mathbf{y}_s)\right]$$
$$= -\frac{1}{m}\left\{m\log(c_1) - m\log(\sigma_v) + \log[\pi(\boldsymbol{\lambda})] + \sum_{i=1}^{m}\log\left[\sum_{t=1}^{T}\psi_{it}(\boldsymbol{\lambda})\right]\right\}; \tag{4.62}$$

and

$$\psi_{it}(\boldsymbol{\lambda}) = \frac{w_t \exp\left[y_i(\mathbf{x}_i'\boldsymbol{\beta} + v_t) - \dfrac{c_0^{1/(2\varphi_v)}}{\sigma_v^{1/\varphi_v}}\left|v_t\right|^{1/\varphi_v} + v_t^2\right]}{\left[1+\exp(\mathbf{x}_i'\boldsymbol{\beta} + v_t)\right]^{n_i}}. \tag{4.63}$$

We next apply both first- and second-order Laplace approximations to estimate the posterior mean and variance of $G(\boldsymbol{\lambda})$.

First-order Laplace approximation:

Following Kass and Steffey (1989, formulas 3.5 and 3.6), we can get the first-order Laplace's approximation for the posterior mean and variance of $G(\boldsymbol{\lambda})$ as:

$$E\left[G(\boldsymbol{\lambda})\,|\,\mathbf{y}_s\right] = G(\hat{\boldsymbol{\lambda}}) + O(m^{-1}); \tag{4.64}$$

$$V\left[G(\boldsymbol{\lambda})\,|\,\mathbf{y}_s\right] = [DG(\hat{\boldsymbol{\lambda}})]'\tilde{\boldsymbol{\Sigma}}[DG(\hat{\boldsymbol{\lambda}})] + O(m^{-2}); \tag{4.65}$$

where $\hat{\boldsymbol{\lambda}}$ is the posterior mode and $\tilde{\boldsymbol{\Sigma}}$ is the inverse of the Hessian matrix of $mh(\boldsymbol{\lambda})$ evaluated at $\hat{\boldsymbol{\lambda}}$: $\tilde{\boldsymbol{\Sigma}} = \{m[D^2h(\hat{\boldsymbol{\lambda}})]\}^{-1}$.

The posterior mode $\hat{\boldsymbol{\lambda}}$ is the value which maximizes the function $h(\boldsymbol{\lambda})$ defined by (4.62). A numerical method such as the Newton-Raphson or EM algorithm is needed to find $\hat{\boldsymbol{\lambda}}$.

Second-order Laplace approximation:

Assume $G(\boldsymbol{\lambda})$ is a positive function**.** We use the fully exponential form to obtain the second-order Laplace's approximation. Let $\hat{L}(\boldsymbol{\lambda}) = \log\left[L(\boldsymbol{\lambda})\pi(\boldsymbol{\lambda})\right] = \log\left[f(\boldsymbol{\lambda} \mid \mathbf{y}_S)\right]$ and $\hat{L}^*(\boldsymbol{\lambda}) = \log\left[G(\boldsymbol{\lambda})\right] + \hat{L}(\boldsymbol{\lambda})$, where $L(\boldsymbol{\lambda})$ is the likelihood function. Assume that $\hat{\boldsymbol{\lambda}}^*$ is the point that maximizes $\hat{L}^*(\boldsymbol{\lambda})$. Let $\tilde{\boldsymbol{\Sigma}}^* = [-D^2\hat{L}^*(\hat{\boldsymbol{\lambda}}^*)]^{-1}$. $\hat{\boldsymbol{\lambda}}$ and $\tilde{\boldsymbol{\Sigma}}$ are defined in (4.64) and (4.65). Then following Tierney and Kadane (1986) and Kass and Steffey (1989), we have

$$E\left[G(\boldsymbol{\lambda}) \mid \mathbf{y}_S\right] = \left[\frac{\det(\tilde{\boldsymbol{\Sigma}}^*)}{\det(\tilde{\boldsymbol{\Sigma}})}\right]^{1/2} \exp\left[\hat{L}^*(\hat{\boldsymbol{\lambda}}^*) - \hat{L}(\hat{\boldsymbol{\lambda}})\right] + O(m^{-2}); \tag{4.66}$$

$$V\left[G(\boldsymbol{\lambda}) \mid \mathbf{y}_S\right] = E\left[G^2(\boldsymbol{\lambda}) \mid \mathbf{y}_S\right] - E^2\left[G(\boldsymbol{\lambda}) \mid \mathbf{y}_S\right]; \tag{4.67}$$

where $E\left[G^2(\boldsymbol{\lambda}) \mid \mathbf{y}_S\right]$ can be approximated using the same approach as that used to approximate $E\left[G(\boldsymbol{\lambda}) \mid \mathbf{y}_S\right]$. For non-positive $G(\boldsymbol{\lambda})$, one can follow the approach discussed by Tierney, Kass and Kadane (1989). We only consider the case when $G(\boldsymbol{\lambda}) > 0$.

As seen from Section 4.4.1, for the mixed logistic model defined by (4.1) and (4.2), the terms of $E\left[b(\theta_i) \mid \mathbf{y}_S\right]$ and $V\left[b(\theta_i) \mid \mathbf{y}_S\right]$ cannot be written as smooth functions of $\boldsymbol{\lambda}$ analytically. However, given the posterior mode $\hat{\boldsymbol{\lambda}}$ and $\tilde{\boldsymbol{\Sigma}}$ as defined in (4.64) and (4.65), we can obtain the first-order approximation to the posterior mean and variance of $b(\theta_i)$ following Kass and Steffey (1989):

$$E\left[b(\theta_i) \mid \mathbf{y}_S\right] = E\left[b(\theta_i) \mid \mathbf{y}_S, \hat{\boldsymbol{\lambda}}\right]\left[1 + O(m^{-1})\right]; \tag{4.68}$$

$$V\left[b(\theta_i)\,|\,\mathbf{y}_s\right]=\left\{V\left[b(\theta_i)\,|\,\mathbf{y}_s,\hat{\boldsymbol{\lambda}}\right]+\sum_{j,h}\tilde{\sigma}_{jh}\hat{\delta}_j\hat{\delta}_h\right\}\left[1+O(m^{-1})\right]; \tag{4.69}$$

where $\tilde{\sigma}_{jh}$ is the (j,h) component of $\tilde{\boldsymbol{\Sigma}}$ and $\hat{\delta}_j=\left(\partial/\partial\lambda_j\right)E\left[b(\theta_i)\,|\,\mathbf{y}_s,\boldsymbol{\lambda}\right]_{\boldsymbol{\lambda}=\hat{\boldsymbol{\lambda}}}$.

The two terms $E\left[b(\theta_i)\,|\,\mathbf{y}_s,\hat{\boldsymbol{\lambda}}\right]$ and $V\left[b(\theta_i)\,|\,\mathbf{y}_s,\hat{\boldsymbol{\lambda}}\right]$ can be obtained using exactly the same approaches as we discussed in Section 4.4.1.

If $b(\theta_i)=\theta_i$, we can obtain the posterior mean and variance of θ_i using results (4.68) and (4.69). The posterior mean and variance of the finite population mean P_i can also be obtained because they are functions of the posterior mean and variance of θ_i (see Section 3.5 of Chapter 3). Before results of (4.68) and (4.69) can be applied, an essential step is to find the posterior mode $\hat{\boldsymbol{\lambda}}$.

4.5.2 Estimation of the Posterior Mode $\hat{\lambda}$

In Bayesian analysis, $\pi(\boldsymbol{\beta})\propto 1$ is often assumed. Assume $\pi(\boldsymbol{\lambda})\propto\pi(\sigma_v,\varphi_v)=\pi(\sigma_v)\pi(\varphi_v)$. We consider the following uniform prior distribution for σ_v and φ_v:

$\pi(\sigma_v)\propto U(0,K)$, where K is a known large positive number; and

$\pi(\varphi_v)\propto U(0,1)$.

Therefore the log-likelihood of $\boldsymbol{\lambda}$ given $\mathbf{y}_s$ is:

$$\log\left[f(\boldsymbol{\lambda}\mid \mathbf{y}_s)\right] \propto m\log(c_1) - m\log(\sigma_v) + \sum_{i=1}^{m}\log\left[\sum_{t=1}^{T}\psi_{it}(\boldsymbol{\lambda})\right]. \tag{4.70}$$

Kass and Steffey (1989) pointed out that the transformation of σ_v to $\exp(-\tau_v/2)$ is generally preferable in numerical work. Using that transformation for σ_v here, we obtain the log-likelihood of $(\boldsymbol{\beta}, \tau_v, \varphi_v)$ given $\mathbf{y}_s$ as:

$$\begin{aligned} L(\boldsymbol{\beta}, \tau_v, \varphi_v) &= \log\left[f(\boldsymbol{\beta}, \tau_v, \varphi_v \mid \mathbf{y}_s)\right] \\ &\propto m\log(c_1) + \frac{m\tau_v}{2} + \sum_{i=1}^{m}\log\left[\sum_{t=1}^{T}\psi_{it}^{*}(\boldsymbol{\beta}, \tau_v, \varphi_v)\right], \end{aligned} \tag{4.71}$$

where

$$\psi_{it}^{*}(\boldsymbol{\beta}, \tau_v, \varphi_v) = \frac{w_t \exp\left[y_i(\mathbf{x}_i'\boldsymbol{\beta} + v_t) - \dfrac{c_0^{1/(2\varphi_v)}|v_t|^{1/\varphi_v}}{\exp(-\tau_v/2\varphi_v)} + v_t^2\right]}{\left[1+\exp(\mathbf{x}_i'\boldsymbol{\beta} + v_t)\right]^{n_i}}. \tag{4.72}$$

To find the posterior mode $\hat{\boldsymbol{\lambda}}$, we need to obtain $\hat{\boldsymbol{\lambda}}^{*} = (\hat{\boldsymbol{\beta}}, \hat{\tau}_v, \hat{\varphi}_v)$ with respect to the log-likelihood $L(\boldsymbol{\beta}, \tau_v, \varphi_v)$ first. To find $\hat{\boldsymbol{\lambda}}^{*}$, we need to solve the equation $DL(\boldsymbol{\beta}, \tau_v, \varphi_v) = 0$, that is, $\hat{\boldsymbol{\lambda}}^{*}$ is the solution to the following set of equations:

$$
\begin{cases}
\dfrac{\partial L}{\partial \boldsymbol{\beta}} = \displaystyle\sum_{i=1}^{m}\left[\frac{\displaystyle\sum_{t=1}^{T}\frac{\partial \psi_{it}^{*}}{\partial \boldsymbol{\beta}}}{\displaystyle\sum_{t=1}^{T}\psi_{it}^{*}(\boldsymbol{\beta},\tau_v,\varphi_v)}\right] = 0 \\
\dfrac{\partial L}{\partial \tau_v} = \dfrac{m}{2} + \displaystyle\sum_{i=1}^{m}\left[\frac{\displaystyle\sum_{t=1}^{T}\frac{\partial \psi_{it}^{*}}{\partial \tau_v}}{\displaystyle\sum_{t=1}^{T}\psi_{it}^{*}(\boldsymbol{\beta},\tau_v,\varphi_v)}\right] = 0 \\
\dfrac{\partial L}{\partial \varphi_v} = \dfrac{m}{c_1}\dfrac{\partial c_1}{\partial \varphi_v} + \displaystyle\sum_{i=1}^{m}\left[\frac{\displaystyle\sum_{t=1}^{T}\frac{\partial \psi_{it}^{*}}{\partial \varphi_v}}{\displaystyle\sum_{t=1}^{T}\psi_{it}^{*}(\boldsymbol{\beta},\tau_v,\varphi_v)}\right] = 0
\end{cases}
\tag{4.73}
$$

where ∂ denotes the partial derivative. Numerical methods are needed to solve the above equations.

Once $\hat{\boldsymbol{\lambda}}^{*}$ is obtained, a transformation of $\hat{\tau}_v$ back to the original scale produces the posterior mode $\hat{\boldsymbol{\lambda}}$. That is, the posterior mode of $\boldsymbol{\lambda}$ is $\hat{\boldsymbol{\lambda}} = \left(\hat{\boldsymbol{\beta}}, \hat{\sigma}_v, \hat{\varphi}_v\right)$, where $\hat{\sigma}_v = \exp(-\hat{\tau}_v / 2)$.

4.6 Concluding Remarks

We have investigated different approximate methods in making Bayesian inference for our proposed model as alternatives to the MCMC technique. Because Gauss-Hermite Quadrature method uses purely numerical approximation, we used it as a standard method to compare among other approximate methods. Laplace approximations offer simple interpretation of the proposed Bayesian methodology.

However, the results from the empirical study for the simple case when all the hyperparameters are known demonstrated that the precision of the Laplace's method depends on the sample size and it does not work well for small samples. The study also indicates that the MCMC is a competitive method to use for Bayesian inference. Since the small area sample sizes are usually small in practice, we do not recommend Laplace's method for the small area estimation problem.

Chapter 5: Empirical Best Prediction of Small-Area Proportions

5.1 Introduction

As reviewed in Chapter 1, there are primarily two different approaches for making inferences using mixed models: i) the classical prediction approach like the empirical best prediction approach; and ii) the hierarchical Bayesian approach. To estimate small-area proportions, we have explored the hierarchical Bayesian approach for inferences using the proposed *Bernoulli-Logit-EP* model in Chapters 3 and 4. In this chapter, we study the empirical best prediction approach for inference using the same model.

As an alternative to the hierarchical Bayesian approach, the empirical best prediction approach has been frequently used for estimating small-area proportions based on logistic regression models with mixed effects. Several applications in this context were reviewed in Chapter 1 (e.g., Dempster and Tomberlin, 1980; MacGibbon and Tomberlin, 1989; Farrell et al., 1997a, b; Jiang and Lahiri, 2001).

Jiang and Lahiri (2001) developed the Taylor linearization method for binary data using mixed logistic model. In this chapter, we develop the Jiang-Lahiri type frequentist alternative to the hierarchical Bayesian methods. Let y_{ik} denote the binary characteristic of interest associated with the *k*-th unit in the *i*-th area ($k = 1,...,N_i;\ i = 1,...,m$). Let $\mathbf{x}_i = (x_{i1},...,x_{ip})' \in R^P$ be a vector of p known auxiliary variables. Suppose that n_i units are chosen from the N_i population units in area i.

The goal is to estimate the small area proportions $P_i = \Sigma_{k=1}^{N_i} y_{ik} / N_i$, $i = 1,...,m,$ using the sample data.

We organize this chapter as follows: we first represent the mixed model in Section 5.2. In Section 5.3, we present the best predictor (BP) and empirical best predictor (EBP) of the random effect v_i. Section 5.4 studies the mean squared error (MSE) of the EBP of v_i. In Section 5.5, we extend the results of Section 5.3 to predict θ_i. Section 5.6 develops the MSE of the EBP of the mixed effect using a parametric Bootstrap approach. We then study the relationship between the MSE of EBP and HB in Section 5.7. The chapter finishes with some concluding remarks in Section 5.8.

5.2 Small Area Model

In order to estimate the finite small area proportions P_i, $i = 1,...,m$, we consider the model that we studied in Chapter 3 and 4, namely *Bernoulli-Logit-EP*:

$$\text{Level 1: } y_{ik} \mid \theta_i \overset{ind}{\sim} \text{Bernoulli}(\theta_i), \qquad k = 1,...,n_i,\ i = 1,...,m; \tag{5.1}$$

$$\text{Level 2: } \text{logit}(\theta_i) = \mathbf{x}_i'\boldsymbol{\beta} + v_i \text{ , where } v_i \overset{iid}{\sim} EP(0, \sigma, \varphi),\ i = 1,...,m. \tag{5.2}$$

The density function of v_i is defined as:

$$f_{EP}(v_i) = \frac{c_1}{\sigma_v} \exp\left\{ -\left| \frac{\sqrt{c_0} v_i}{\sigma_v} \right|^{1/\varphi_v} \right\}, \tag{5.3}$$

where $\sigma_v \in R^+$, $\varphi_v \in (0,1]$, $c_0 = \dfrac{\Gamma(3\varphi_v)}{\Gamma(\varphi_v)}$, $c_1 = \dfrac{\sqrt{c_0}}{2\varphi_v \Gamma(\varphi_v)}$.

As noted earlier, for the special case $\varphi_v = 0.5$, the random effects v_i are i.i.d. normal, and the proposed model reduces to the mixed logistic regression model, which has been studied in the literature (e.g., see Jiang and Lahiri, 2001).

Based on assumption (5.2), θ_i can be expressed as:

$$\theta_i = \frac{\exp\left(\mathbf{x}_i'\boldsymbol{\beta} + v_i\right)}{1 + \exp\left(\mathbf{x}_i'\boldsymbol{\beta} + v_i\right)}. \tag{5.4}$$

We will show how to make inference for v_i first and then for the parameter θ_i.

5.3 The BP and EBP of v_i

Let $\boldsymbol{\lambda} = \left(\boldsymbol{\beta}, \sigma_v, \varphi_v\right)'$ denote the model parameters and $\boldsymbol{\lambda}_0 = \left(\boldsymbol{\beta}_0, \sigma_0, \varphi_0\right)'$ denote the true value of $\boldsymbol{\lambda}$. Assume that $\boldsymbol{\lambda}$ is known. The posterior distribution of the random effects v_i can be obtained as below:

$$\begin{aligned} f(v_i \mid \mathbf{y}_s) &\propto f(v_i \mid y_i) \\ &= \left[\prod_{j=1}^{n_i} f(y_{ij} \mid v_i)\right] f(v_i) \\ &= \frac{\exp(y_i \mathbf{x}_i'\boldsymbol{\beta} + y_i v_i)}{\left[1 + \exp(\mathbf{x}_i'\boldsymbol{\beta} + v_i)\right]^{n_i}} \frac{c_1}{\sigma_v} \exp\left(-\left|\frac{\sqrt{c_0} v_i}{\sigma_v}\right|^{1/\varphi_v}\right). \end{aligned}$$

Therefore, the posterior mean of v_i, also called the best predictor of v_i, is:

$$
\begin{aligned}
E(v_i \mid \mathbf{y}_s) &= \frac{\int v_i f(v_i \mid \mathbf{y}_s) dv_i}{\int f(v_i \mid \mathbf{y}_s) dv_i} \\
&= \frac{\int v_i \dfrac{\exp(y_i \mathbf{x}_i'\boldsymbol{\beta}_0 + y_i v_i)}{\left[1+\exp(\mathbf{x}_i'\boldsymbol{\beta}_0 + v_i)\right]^{n_i}} \dfrac{c_1}{\sigma_0} \exp\left(-\left| \dfrac{\sqrt{c_0} v_i}{\sigma_0} \right|^{1/\varphi_0} \right) dv_i}{\int \dfrac{\exp(y_i \mathbf{x}_i'\boldsymbol{\beta}_0 + y_i v_i)}{\left[1+\exp(\mathbf{x}_i'\boldsymbol{\beta}_0 + v_i)\right]^{n_i}} \dfrac{c_1}{\sigma_0} \exp\left(-\left| \dfrac{\sqrt{c_0} v_i}{\sigma_0} \right|^{1/\varphi_0} \right) dv_i} \\
&= \frac{\int v_i \dfrac{\exp(y_i v_i)}{\left[1+\exp(\mathbf{x}_i'\boldsymbol{\beta}_0 + v_i)\right]^{n_i}} \exp\left(-\left| \dfrac{\sqrt{c_0} v_i}{\sigma_0} \right|^{1/\varphi_0} \right) dv_i}{\int \dfrac{\exp(y_i v_i)}{\left[1+\exp(\mathbf{x}_i'\boldsymbol{\beta}_0 + v_i)\right]^{n_i}} \exp\left(-\left| \dfrac{\sqrt{c_0} v_i}{\sigma_0} \right|^{1/\varphi_0} \right) dv_i}.
\end{aligned}
$$

Letting $v_i = \sigma_0 \xi$, where $\xi \sim EP(0,\ 1,\ \varphi_0)$, we can rewrite $E(v_i \mid \mathbf{y}_s)$ as:

$$
\begin{aligned}
E(v_i \mid \mathbf{y}_s) &= \sigma_0 \frac{\int \xi \dfrac{\exp(\sigma_0 y_i \xi)}{\left[1+\exp(\mathbf{x}_i'\boldsymbol{\beta}_0 + \sigma_0 \xi)\right]^{n_i}} \exp\left(-\left| \sqrt{c_0}\xi \right|^{1/\varphi_0} \right) d\xi}{\int \dfrac{\exp(\sigma_0 y_i \xi)}{\left[1+\exp(\mathbf{x}_i'\boldsymbol{\beta}_0 + \sigma_0 \xi)\right]^{n_i}} \exp\left(-\left| \sqrt{c_0}\xi \right|^{1/\varphi_0} \right) d\xi} \\
&= \sigma_0 \frac{E\left\{ \xi \exp\left[\phi_i(y_i, \sigma_0 \xi, \boldsymbol{\beta}_0) \right] \right\}}{E\left\{ \exp\left[\phi_i(y_i, \sigma_0 \xi, \boldsymbol{\beta}_0) \right] \right\}} \equiv \psi_i(y_i, \lambda_0),
\end{aligned}
\tag{5.5}
$$

where $\phi_i(y_i, \sigma_0 \xi, \boldsymbol{\beta}_0) = \sigma_0 y_i \xi - n_i \log\left[1+\exp(\mathbf{x}_i'\boldsymbol{\beta}_0 + \sigma_0 \xi)\right]$.

When $\boldsymbol{\lambda}_0$ is given, $\psi_i(y_i, \boldsymbol{\lambda}_0)$ can be computed using a numerical integration method, such as Gauss-Hermite Quadrature or a Laplace approximation. We prefer the GHQ method since the sample sizes n_i are small. The quality $\psi_i(y_i, \boldsymbol{\lambda}_0)$ is called the best predictor (BP) of v_i, i.e., $\hat{v}_i^{BP} = \psi_i(y_i, \boldsymbol{\lambda}_0)$.

Since λ_0 is unknown in practice, $\psi_i(y_i, \boldsymbol{\lambda}_0)$ is not computable. It is customary to replace $\boldsymbol{\lambda}_0$ by a consistent estimator $\tilde{\boldsymbol{\lambda}}$ in $\psi_i(y_i, \boldsymbol{\lambda}_0)$. The resulting estimator is called empirical best predictor (EBP) of v_i:

$$\hat{v}_i^{EBP} = \psi_i(y_i, \tilde{\boldsymbol{\lambda}}).$$

The maximum likelihood (ML) approach can be used to estimate $\boldsymbol{\lambda}$.

5.4 The MSE of the EBP of v_i

To derive the MSE of the EBP of v_i, we follow the approach used by Jiang and Lahiri (2001).

According to the definition, the MSE of $\hat{v}_i^{EBP}$ is:

$$\begin{aligned} MSE(\hat{v}_i^{EBP}) &= E(\hat{v}_i^{EBP} - v_i)^2 \\ &= E\left[\hat{v}_i^{EBP} - E(v_i \mid \mathbf{y}_s)\right]^2 + E\left[E(v_i \mid \mathbf{y}_s) - v_i\right]^2. \end{aligned} \tag{5.6}$$

The second term on the right side of (5.6) has a closed form:

$$\begin{aligned} E\left[E(v_i \mid \mathbf{y}_s) - v_i\right]^2 &= E(v_i^2) - E\left[E(v_i \mid \mathbf{y}_s)\right]^2 \\ &= \sigma_0^2 - E\left[E(v_i \mid \mathbf{y}_s)\right]^2, \end{aligned}$$

and

$$E\left[E(v_i \mid \mathbf{y}_s)\right]^2 = E\left[\psi_i(y_i, \boldsymbol{\lambda}_0)\right]^2 = \sum_{k=0}^{n_i} \psi_i^2(k, \boldsymbol{\lambda}_0) p_i(k, \boldsymbol{\lambda}_0) \equiv b_i(\boldsymbol{\lambda}_0), \tag{5.7}$$

where

$$
\begin{aligned}
p_i(k,\boldsymbol{\lambda}_0) &= P(y_i = k) \\
&= E\left[I(y_i = k)\right] \\
&= E\left\{E\left[I(y_i = k)\,|\,v_i\right]\right\} \\
&= E\left\{E\frac{\exp(k\mathbf{x}_i'\boldsymbol{\beta}_0 + kv_i)}{\left[1+\exp(\mathbf{x}_i'\boldsymbol{\beta}_0 + v_i)\right]^{n_i-k}}\right\} \\
&= E\left\{\exp(k\mathbf{x}_i'\boldsymbol{\beta}_0)E\exp\left[\phi_i(k,\sigma_0\xi,\boldsymbol{\beta}_0)\right]\right\} \\
&= \sum_{z\in S(n_i,k)} \exp\left(\mathbf{x}_i'\boldsymbol{\beta}_0\sum_{j=1}^{n_i} z_j\right) E\exp\left[\phi_i(z,\sigma_0\xi,\boldsymbol{\beta}_0)\right],
\end{aligned}
$$

with $S(n_i,k) = \left\{z = (z_1,...,z_{n_i}) \in \{0,1\}^{n_i},\ z_\bullet = z_1 + ... + z_{n_i} = k\right\}$.

For the first term on the right side of (5.6), we use a Taylor series expansion:

$$
\begin{aligned}
\hat{v}_i^{EBP} - E(v_i\,|\,\mathbf{y}_s) &= \psi_i(y_i,\tilde{\boldsymbol{\lambda}}) - \psi_i(y_i,\boldsymbol{\lambda}_0) \\
&= \left[\frac{\partial}{\partial\boldsymbol{\lambda}}\psi_i(y_i,\boldsymbol{\lambda}_0)\right]'(\tilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0) \qquad (5.8) \\
&\quad + \frac{1}{2}(\tilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0)'\left[\frac{\partial^2}{\partial\boldsymbol{\lambda}^2}\psi_i(y_i,\boldsymbol{\lambda}_0)\right](\tilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0) + o\left(|\,\tilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0\,|^2\right).
\end{aligned}
$$

Suppose that:

$$
|\,\tilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0\,| = O_p(1/\sqrt{n}), \qquad (5.9)
$$

where $n = \Sigma_{i=1}^m n_i$ is the total sample size. When $\tilde{\boldsymbol{\lambda}}$ is a ML estimator of $\boldsymbol{\lambda}$, equation (5.9) still holds (Bradley and Gart, 1962). It is expected that

$$
E\left[\hat{v}_i^{EBP} - E(v_i\,|\,\mathbf{y}_s)\right]^2 = \frac{1}{n}E\left\{\left[\frac{\partial}{\partial\boldsymbol{\lambda}}\psi_i(y_i,\boldsymbol{\lambda}_0)\right]'\sqrt{n}(\tilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0)\right\}^2 + o\left(\frac{1}{n}\right). \qquad (5.10)
$$

Now, assume $\tilde{\boldsymbol{\lambda}} = \tilde{\boldsymbol{\lambda}}_{i-}$, an estimator of $\boldsymbol{\lambda}$ based on y_{i-}, where y_{i-} denotes the observed data after deleting the ith area. Write $\hat{v}_{i-} = \psi_i(y_i, \tilde{\boldsymbol{\lambda}}_{i-})$. Then by independence of y_i and y_{i-}, we have

$$
\begin{aligned}
& E\left\{\left[\frac{\partial}{\partial \boldsymbol{\lambda}}\psi_i(y_i, \boldsymbol{\lambda}_0)\right]' \sqrt{n}(\tilde{\boldsymbol{\lambda}}_{i-} - \boldsymbol{\lambda}_0)\right\}^2 \\
&= E\left\{E\left\{\left\{\left[\frac{\partial}{\partial \boldsymbol{\lambda}}\psi_i(y_i, \boldsymbol{\lambda}_0)\right]' \sqrt{n}(\tilde{\boldsymbol{\lambda}}_{i-} - \boldsymbol{\lambda}_0)\right\}^2 \Big| y_i = k\right\}\right\} \\
&= E\left\{\left[\frac{\partial}{\partial \boldsymbol{\lambda}}\psi_i(k, \boldsymbol{\lambda}_0)\right]' V_i(\boldsymbol{\lambda}_0)\left[\frac{\partial}{\partial \boldsymbol{\lambda}}\psi_i(k, \boldsymbol{\lambda}_0)\right]\right\} \\
&= \sum_{k=0}^{n_i}\left[\frac{\partial}{\partial \lambda}\psi_i(k, \boldsymbol{\lambda}_0)\right]' V_i(\boldsymbol{\lambda}_0)\left[\frac{\partial}{\partial \lambda}\psi_i(k, \boldsymbol{\lambda}_0)\right] p_i(k, \boldsymbol{\lambda}_0) \equiv a_i(\boldsymbol{\lambda}_0),
\end{aligned}
\tag{5.11}
$$

where $V_i(\boldsymbol{\lambda}_0) = nE(\tilde{\boldsymbol{\lambda}}_{i-} - \boldsymbol{\lambda}_0)(\tilde{\boldsymbol{\lambda}}_{i-} - \boldsymbol{\lambda}_0)'$.

Combing (5.6)-(5.8), (5.10) and (5.11), we obtain

$$
MSE(\hat{v}_{i-}^{EBP}) = \sigma_0^2 - b_i(\boldsymbol{\lambda}_0) + (1/n)a_i(\boldsymbol{\lambda}_0) + o(1/n), \tag{5.12}
$$

where $b_i(\boldsymbol{\lambda}_0)$ is defined in (5.7) and $a_i(\boldsymbol{\lambda}_0)$ is defined in (5.11).

The result (5.12) is based on the assumption that $\boldsymbol{\lambda} = \tilde{\boldsymbol{\lambda}}_{i-}$. Next we want to evaluate how close $MSE(\hat{v}_{i-}^{EBP})$ is to $MSE(\hat{v}_i^{EBP})$.

$$
\begin{aligned}
MSE(\hat{v}_i^{EBP}) &= MSE(\hat{v}_{i-}^{EBP}) + 2E(\hat{v}_i^{EBP} - \hat{v}_{i-}^{EBP})(\hat{v}_{i-}^{EBP} - v_i) \\
&\quad + E(\hat{v}_i^{EBP} - \hat{v}_{i-}^{EBP})^2 \\
&= MSE(\hat{v}_{i-}^{EBP}) + r_i,
\end{aligned}
\tag{5.13}
$$

where $r_i = 2E(\hat{v}_i^{EBP} - \hat{v}_{i-}^{EBP})(\hat{v}_{i-}^{EBP} - v_i) + E(\hat{v}_i^{EBP} - \hat{v}_{i-}^{EBP})^2$.

Based on (5.9), Jiang and Lahiri (2001) showed that it is reasonable to assume that

$$|\tilde{\boldsymbol{\lambda}} - \tilde{\boldsymbol{\lambda}}_{i-}| = O_p(1/\sqrt{n}). \tag{5.14}$$

It follows from (5.9), (5.14), and the Taylor series expansion that $r_i = o(1/n)$. Note that $E(\hat{v}_i^{EBP} - \hat{v}_{i-}^{EBP})(\hat{v}_{i-}^{EBP} - v_i) = E(\hat{v}_i^{EBP} - \hat{v}_{i-}^{EBP})(\hat{v}_{i-}^{EBP} - \hat{v}_i^{BP})$. Therefore, by (5.13) and (5.14), we have

$$\begin{aligned} MSE(\hat{v}_i^{EBP}) &= MSE(\hat{v}_{i-}^{EBP}) + o(1/n) \\ &= \sigma_0^2 - b_i(\boldsymbol{\lambda}_0) + (1/n)c_i(\boldsymbol{\lambda}_0) + o(1/n), \end{aligned} \tag{5.15}$$

where $c_i(\boldsymbol{\lambda}_0)$ is the same as $a_i(\boldsymbol{\lambda}_0)$ except that $V_i(\boldsymbol{\lambda}_0)$ is replaced by $V(\boldsymbol{\lambda}_0) = nE(\tilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0)(\tilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0)'$.

5.5 The BP and EBP of θ_i

Since θ_i is a function of $\boldsymbol{\beta}$ and v_i from (5.4), the development of inferences for v_i in Section 5.3 and 5.4 can be extended in a parallel way to make inferences for θ_i, and then for P_i.

Let $\theta_i = h_i(\boldsymbol{\beta}, v_i)$, where the function h is defined by (5.4). The best predictor for θ_i is:

$$\theta_i^{BP} = E(\theta_i \mid \mathbf{y}_s, \boldsymbol{\lambda}) = \frac{E\left\{h_i(\boldsymbol{\beta}_0, \sigma_0\xi)\exp\left[\phi_i(y_i, \sigma_0\xi, \boldsymbol{\beta}_0)\right]\right\}}{E\left\{\exp\left[\phi_i(y_i, \sigma_0\xi, \boldsymbol{\beta}_0)\right]\right\}} \equiv \tilde{\psi}_i(y_i, \boldsymbol{\lambda}_0). \tag{5.16}$$

As shown in Chapter 4, the integrals involved on the right-hand side of (5.16) can be approximated using numerical methods or Laplace's method, and again numerical methods are preferred here because of small sample sizes.

Replace $\boldsymbol{\lambda}_0$ in (5.16) by a consistent estimator $\tilde{\boldsymbol{\lambda}}$, we get the EBP for P_i, given by

$$\theta_i^{EBP} = \tilde{\psi}_i(y_i, \tilde{\boldsymbol{\lambda}}). \tag{5.17}$$

When the MLE of $\boldsymbol{\lambda}$ is used to obtain the EBP, the EBP of θ_i is very close to the HB of θ_i approximated using the first-order Laplace's method (see Chapter 4).

5.6 The MSE of the EBP of θ_i

The MSE of θ_i^{EBP} is:

$$MSE(\theta_i^{EBP}) = MSE(\theta_i^{BP}) + E(\theta_i^{EBP} - \theta_i^{BP})^2. \tag{5.18}$$

The first term on the right side of (5.18) is the MSE of the BP, while the second term is the approximate mean squared of the EBP to the BP, which measures the uncertainty due to the estimation of $\boldsymbol{\lambda}$. Furthermore, we have

$$\begin{aligned} MSE(\theta_i^{BP}) &= E\left[E(\theta_i \mid \mathbf{y}_s, \boldsymbol{\lambda}) - \theta_i\right]^2 \\ &= E\left[E^2(\theta_i \mid \mathbf{y}_s, \boldsymbol{\lambda}) - 2\theta_i E(\theta_i \mid \mathbf{y}_s, \boldsymbol{\lambda}) + \theta_i^2\right] \\ &= E(\theta_i^2) - E\left[E^2(\theta_i \mid \mathbf{y}_s, \boldsymbol{\lambda})\right] \\ &= E(\theta_i^2) - E(\theta_i^{BP})^2 \\ &= Eh_i^2(\boldsymbol{\beta}_0, \sigma_0 \xi) - \sum_{k=0}^{n_i} \tilde{\psi}_i^2(k, \boldsymbol{\lambda}_0) p_i(k, \boldsymbol{\lambda}_0) \equiv \tilde{d}_i(\boldsymbol{\lambda}_0), \end{aligned} \tag{5.19}$$

where $\tilde{\psi}_i(k,\boldsymbol{\lambda}_0)$ is defined in (5.16).

If $\theta_i^{EBP} = \theta_{i-}^{EBP}$, an estimator of P_i based on y_{i-}, then

$$E(\theta_i^{EBP} - \theta_i^{BP})^2 = (1/n)\tilde{a}_i(\boldsymbol{\lambda}_0) + o(1/n),$$

where $\tilde{a}_i(\boldsymbol{\lambda}_0)$ is $a_i(\boldsymbol{\lambda}_0)$ as defined in (5.11) except that $\psi_i(k,\boldsymbol{\lambda}_0)$ in $a_i(\boldsymbol{\lambda}_0)$ is replaced by $\tilde{\psi}_i(k,\boldsymbol{\lambda}_0)$.

Thus with $\theta_{i-}^{EBP} = \tilde{\psi}_i(y_i, \tilde{\boldsymbol{\lambda}}_{i-})$,

$$MSE(\theta_{i-}^{EBP}) = \tilde{d}_i(\boldsymbol{\lambda}_0) + \tilde{a}_i(\boldsymbol{\lambda}_0)/n + o(1/n), \tag{5.20}$$

where $\tilde{d}_i(\boldsymbol{\lambda}_0)$ is defined in (5.19).

Also one may replace θ_{i-}^{EBP} by θ_i^{EBP}, an estimator of P_i based on all the data, may still obtain

$$\begin{aligned} MSE(\theta_i^{EBP}) &= MSE(\theta_{i-}^{EBP}) + o(1/n) \\ &= \tilde{d}_i(\boldsymbol{\lambda}_0) + \tilde{c}_i(\boldsymbol{\lambda}_0)/n + o(1/n), \end{aligned} \tag{5.21}$$

where $\tilde{d}_i(\boldsymbol{\lambda}_0) = Eh_i^2(\boldsymbol{\beta}_0, \sigma_0\xi) - \sum_{k=0}^{n_i} \tilde{\psi}_i^2(k,\boldsymbol{\lambda}_0) p_i(k,\boldsymbol{\lambda}_0)$ which is defined in (5.19); and

$\tilde{c}_i(\boldsymbol{\lambda}_0) = \sum_{k=0}^{n_i} \left[\frac{\partial}{\partial\boldsymbol{\lambda}} \tilde{\psi}_i(k,\boldsymbol{\lambda}_0)\right]' V(\boldsymbol{\lambda}_0) \left[\frac{\partial}{\partial\boldsymbol{\lambda}} \tilde{\psi}_i(k,\boldsymbol{\lambda}_0)\right] p_i(k,\boldsymbol{\lambda}_0)$, where $\tilde{c}_i(\boldsymbol{\lambda}_0)$ is $\tilde{a}_i(\boldsymbol{\lambda}_0)$ except that $V_i(\boldsymbol{\lambda}_0)$ in $\tilde{a}_i(\boldsymbol{\lambda}_0)$ is replaced by $V(\boldsymbol{\lambda}_0) = nE(\tilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0)(\tilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0)'$;

5.6.1 Estimate the MSE of the EBP using Taylor Series Linearization

The MSE of θ_i^{EBP} defined by (5.21) involves the true $\boldsymbol{\lambda}_0$ which is unknown in practice. For practical applications, we need to estimate $MSE(\theta_i^{EBP})$ using an

estimator of $\boldsymbol{\lambda}_0$. There are different approaches for estimating $MSE(\theta_i^{EBP})$. In this subsection, we propose a second-order approximation of the MSE of EBP using the Taylor series linearization method.

The Taylor series linearization method has been frequently used to approximate the MSE of an empirical best linear unbiased predictor (EBLUP) in the small area estimation literature. For example, Prasad and Rao (1990) proposed a second-order Taylor series approximation to the MSE of EBLUP for three linear mixed models all with the normality assumption for the distribution of the model effects. Other references can be seen in Kleffe and Rao (1992), Lahiri and Rao (1995), Datta and Lahiri (2000), and Butar and Lahiri (2002). However, the literature on the assessment of the uncertainty of EBP for binary data is limited. The available references include Jiang and Lahiri (2001, 2006b). Jiang and Lahiri (2006a) provides additional references. We develop a second-order approximation approach to measure the uncertainty of the MSE estimate of the EBP based on the proposed model next.

A naïve approach approximates the MSE of θ_i^{EBP} using only the first term $\tilde{d}_i(\boldsymbol{\lambda}_0)$ in (5.21). Replacing $\boldsymbol{\lambda}_0$ by a consistent estimator $\tilde{\boldsymbol{\lambda}}$, a naïve estimator of $MSE(\theta_i^{EBP})$ is as follows:

$$MSE^{naive}(\theta_i^{EBP}) = \tilde{d}_i(\tilde{\boldsymbol{\lambda}}). \tag{5.22}$$

If n is small, this naïve approximation could lead to serious underestimation of the MSE defined by (5.21) for the following two reasons:

i) The second term, $\tilde{c}_i(\boldsymbol{\lambda}_0)/n$, is of order $O(1/n)$. When n is small, we should not ignore any term of order $O(1/n)$.

ii) Replacing the true value $\boldsymbol{\lambda}_0$ by an estimator $\hat{\boldsymbol{\lambda}}$ introduces additional bias of order $O(1/n)$.

An improved approach approximates the MSE of θ_i^{EBP} by incorporating both terms in (5.21) after replacing $\boldsymbol{\lambda}_0$ with a consistent estimator $\tilde{\boldsymbol{\lambda}}$. That is, an improved estimator of $MSE(\theta_i^{EBP})$ is given by:

$$MSE^{IM}(\theta_i^{EBP}) = \tilde{d}_i(\tilde{\boldsymbol{\lambda}}) + \tilde{c}_i(\tilde{\boldsymbol{\lambda}})/n. \tag{5.23}$$

The estimator $MSE^{IM}(\theta_i^{EBP})$ improves the naïve estimator $MSE^{naive}(\theta_i^{EBP})$ by taking account of some terms of order $O(1/n)$. However, this improved MSE estimator still does not account for the additional bias due to the estimation of $\boldsymbol{\lambda}_0$.

Now the question is how to correct the additional bias due to the estimation of $\boldsymbol{\lambda}_0$. To do that, we first need calculate this additional bias.

Using (5.9), the Taylor series expansion of $\tilde{d}_i(\tilde{\boldsymbol{\lambda}})$ around $\boldsymbol{\lambda}_0$ gives:

$$\tilde{d}_i(\tilde{\boldsymbol{\lambda}}) \approx \tilde{d}_i(\boldsymbol{\lambda}_0) + (\tilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0)'[D\tilde{d}_i(\boldsymbol{\lambda}_0)] + \frac{1}{2}\left[(\tilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0)' H_d(\boldsymbol{\lambda}_0)(\tilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0)\right] + o_p(1/n), \tag{5.24}$$

where $D\tilde{d}_i(\boldsymbol{\lambda}_0)$ is the first derivative to function $\tilde{d}_i(\boldsymbol{\lambda}_0)$ with respect to the vector $\boldsymbol{\lambda}_0$, and $H_{d_i}(\boldsymbol{\lambda}_0)$ is the Hessian matrix of $\tilde{d}_i(\boldsymbol{\lambda}_0)$ at value $\boldsymbol{\lambda}_0$.

Since $\left|\tilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0\right| = O_p\left(1/\sqrt{n}\right)$, we write $E(\tilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0) = \frac{1}{\sqrt{n}} B(\boldsymbol{\lambda}_0) + o(1/\sqrt{n})$, where the order of $B(\boldsymbol{\lambda}_0)$ is $O(1)$. In addition, let $E(\tilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0)(\tilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0)' = \frac{1}{n} V(\boldsymbol{\lambda}_0)$. Then

$$E\left[\tilde{d}_i(\tilde{\boldsymbol{\lambda}})\right] = \tilde{d}_i(\boldsymbol{\lambda}_0) + \frac{1}{\sqrt{n}} B(\boldsymbol{\lambda}_0)'[D\tilde{d}_i(\boldsymbol{\lambda}_0)] + \frac{1}{2n} trace\left[V(\boldsymbol{\lambda}_0) H_{d_i}(\boldsymbol{\lambda}_0)\right] + o(1/n). \tag{5.25}$$

Therefore, the additional bias introduced by the first term of $\tilde{d}_i(\tilde{\boldsymbol{\lambda}})$ due to the estimation of $\boldsymbol{\lambda}_0$ is as follows:

$$bias\left[\tilde{d}_i(\tilde{\boldsymbol{\lambda}})\right] \approx \frac{1}{\sqrt{n}} B(\boldsymbol{\lambda}_0)'[D\tilde{d}_i(\boldsymbol{\lambda}_0)] + \frac{1}{2n} trace\left[V(\boldsymbol{\lambda}_0) H_d(\boldsymbol{\lambda}_0)\right], \tag{5.26}$$

which is of order $O(1/n)$.

Similarly, we can get the expectation of $\tilde{c}_i(\tilde{\boldsymbol{\lambda}})$ using Taylor series expansion around $\boldsymbol{\lambda}_0$ as below:

$$E\left\{\tilde{c}_i(\tilde{\boldsymbol{\lambda}})\right\} = \tilde{c}_i(\boldsymbol{\lambda}_0) + \frac{1}{\sqrt{n}} B(\boldsymbol{\lambda}_0)'[D\tilde{c}_i(\boldsymbol{\lambda}_0)] + \frac{1}{2n} trace\left[V(\boldsymbol{\lambda}_0) H_{c_i}(\boldsymbol{\lambda}_0)\right] + o(1/n), \tag{5.27}$$

where $D\tilde{c}_i(\boldsymbol{\lambda}_0)$ is the first derivative of the function $\tilde{c}_i(\boldsymbol{\lambda}_0)$ with respect to the vector $\boldsymbol{\lambda}_0$, and $H_{c_i}(\boldsymbol{\lambda}_0)$ is the Hessian matrix of $\tilde{c}_i(\boldsymbol{\lambda}_0)$ evaluated at $\boldsymbol{\lambda}_0$. It follows from (5.27) that the expectation of $\tilde{c}_i(\tilde{\boldsymbol{\lambda}})/n$ is:

$$E\left\{\tilde{c}_i(\tilde{\boldsymbol{\lambda}})/n\right\} = \tilde{c}_i(\boldsymbol{\lambda}_0)/n + o(1/n), \tag{5.28}$$

which indicates that $\tilde{c}_i(\tilde{\boldsymbol{\lambda}})/n$ is approximately unbiased for $\tilde{c}_i(\boldsymbol{\lambda}_0)/n$ up to the order $O(1/n)$.

Now, from (5.25) and (5.27), after correcting the bias of $\tilde{d}_i(\tilde{\boldsymbol{\lambda}})$ as defined by (5.25), we obtain a second-order unbiased estimator of the MSE of θ_i^{EBP}:

$$MSE^{TS}(\theta_i^{EBP}) = \tilde{d}_i(\tilde{\boldsymbol{\lambda}}) + \tilde{c}_i(\tilde{\boldsymbol{\lambda}})/n - \frac{1}{\sqrt{n}} B(\tilde{\boldsymbol{\lambda}})'[D\tilde{d}_i(\tilde{\boldsymbol{\lambda}})] - \frac{1}{2n} trace\left[V(\tilde{\boldsymbol{\lambda}}) H_{d_i}(\tilde{\boldsymbol{\lambda}})\right], \tag{5.29}$$

which is of order $O(1/n)$.

5.6.2 Estimating the MSE of the EBP using a Parametric Bootstrap

We have developed a second-order MSE approximation $MSE^{TS}(\theta_i^{EBP})$ to measure the uncertainty of θ_i^{EBP} using the Taylor series linearization method. However, there are a few disadvantages associated with this estimator including 1) The MSE estimator could be negative; and 2) the calculation of the MSE estimator requires high order derivatives of a complex function and variance estimation for a consistent estimator of $\boldsymbol{\lambda}$. In order to overcome these disadvantages, we develop a second-order bias corrected computationally simple technique using the two-level parametric bootstrap method following Chatterjee and Lahiri (2008).

Let $\mathbf{y}_{mn_m} = (y_{11},...,y_{1n_1};...;y_{m1},...,y_{mn_m})$ denote all the observed data. Our two-level parametric bootstrap for generating resamples is given below:

1. Resample $\mathbf{y}^*_{mn_m} = (y^*_{11},...,y^*_{1n_1};...;y^*_{m1},...,y^*_{mn_m})$ using the following two-level model:

$$\text{Level 1: } y^*_{ik} \mid \theta_i^* \overset{ind}{\sim} \text{Bernoulli}(\theta_i^*), \quad k = 1,...,n_i,\ i = 1,...,m, \tag{5.30}$$

$$\text{Level 2: } \text{logit}(\theta_i^*) = \mathbf{x}_i'\tilde{\boldsymbol{\beta}} + v_i^*, \tag{5.31}$$

$$\text{where } v_i^* \overset{iid}{\sim} EP(0, \tilde{\sigma}_v, \tilde{\varphi}_v),\ i = 1,...,m, \tag{5.32}$$

and $\tilde{\boldsymbol{\lambda}} = (\tilde{\boldsymbol{\beta}}, \tilde{\sigma}_v, \tilde{\varphi}_v)$ is a consistent estimator obtained using the original sample data $\mathbf{y}_{mn_m}$. The resampling requires that the v_i^* are generated first using (5.32), then the θ_i^* are generated using (5.31), and finally the y_{ik}^*, $k = 1,...,n_i$, $i = 1,...,m$, are generated using (5.30). The expectation of this step, which is conditional on $\mathbf{y}_{mn_m}$, is denoted by E^*.

2. Obtain $\tilde{\boldsymbol{\lambda}}^* = \left(\tilde{\boldsymbol{\beta}}^*, \tilde{\sigma}_v^*, \tilde{\varphi}_v^*\right)$ based on the resamples $\mathbf{y}_{mn_m}^*$ using exactly the same approach as we obtain $\tilde{\boldsymbol{\lambda}}$ from the original data $\mathbf{y}_{mn_m}$.

3. Resample $\mathbf{y}_{mn_m}^{**} = (y_{11}^{**},...,y_{1n_1}^{**};...;y_{m1}^{**},...,y_{mn_m}^{**})$ using the following two level model:

$$\text{Level 1: } y_{ik}^{**} \mid \theta_i^{**} \overset{ind}{\sim} \text{Bernoulli}(\theta_i^{**}); \qquad k = 1,...,n_i;\ i = 1,...,m \tag{5.33}$$

$$\text{Level 2: } \text{logit}(\theta_i^{**}) = \mathbf{x}_i'\tilde{\boldsymbol{\beta}}^* + v_i^{**}, \tag{5.34}$$

$$\text{where } v_i^{**} \overset{iid}{\sim} EP(0, \tilde{\sigma}_v^*, \tilde{\varphi}_v^*),\ i = 1,...,m, \tag{5.35}$$

and $\tilde{\lambda}^* = (\tilde{\boldsymbol{\beta}}^*, \tilde{\sigma}_v^*, \tilde{\varphi}_v^*)$ is obtained at step 2. The resampling requires that the v_i^{**} are generated first using model (5.35), then the θ_i^{**} are generated using model (5.34), and finally the y_{ik}^{**}, $k = 1,...,n_i$, $i = 1,...,m$, are generated using model (5.33). The expectation of this step, which is conditional on $\mathbf{y}_{mn_m}$ and $\mathbf{y}_{mn_m}^*$, is denoted by E^{**}.

4. We define the following statistics:

$\tilde{\boldsymbol{\lambda}}^{**} = \tilde{\boldsymbol{\lambda}}(\mathbf{y}_{mn_m}^{**})$, the consistent estimator of $\boldsymbol{\lambda}$ obtained from $\mathbf{y}_{mn_m}^{**}$;

$$M^{*} = E^{*}\left[\theta_i^{*} - \theta_i^{EBP}(\mathbf{y}_{mn_m}^{*}, \tilde{\boldsymbol{\lambda}}^{*})\right]^2, \tag{5.36}$$

where $\theta_i^{EBP}(\mathbf{y}_{mn_m}^{*}, \tilde{\boldsymbol{\lambda}}^{*})$ denotes the EBP of P_i based on $\mathbf{y}_{mn_m}^{*}$ and $\tilde{\boldsymbol{\lambda}}^{*}$;

$$M^{**} = E^{*}E^{**}\left[\theta_i^{**} - \theta_i^{EBP}(\mathbf{y}_{mn_m}^{**}, \tilde{\boldsymbol{\lambda}}^{**})\right]^2, \tag{5.37}$$

where $\theta_i^{EBP}(\mathbf{y}_{mn_m}^{**}, \tilde{\boldsymbol{\lambda}}^{**})$ denotes the EBP of P_i based on $\mathbf{y}_{mn_m}^{**}$ and $\tilde{\boldsymbol{\lambda}}^{**}$.

Once M^{*} and M^{**} are obtained, following Chatterjee and Lahiri (2008), the following four parametric bootstrap estimators of $MSE(\theta_i^{EBP})$, which are all functions of M^{*} and M^{**}, can be considered:

$$MSE_1^{BOOT}(\theta_i^{EBP}) = (2M^{*} - M^{**})I_{\left(M^{*} > \frac{M^{**}}{2}\right)}; \tag{5.38}$$

$$\begin{aligned} MSE_2^{BOOT}(\theta_i^{EBP}) = {} & (2M^{*} - M^{**})I_{\left(M^{*} \geq M^{**}\right)} \\ & + \left[M^{*}\exp\left(\frac{M^{*}}{M^{**}} - 1\right)\right] I_{\left(M^{*} < M^{**}\right)} \end{aligned}; \tag{5.39}$$

$$\begin{aligned} MSE_3^{BOOT}(\theta_i^{EBP}) = {} & \left\{M^{*} + n^{-1}\tan^{-1}\left[n(M^{*} - M^{**})\right]\right\} I_{\left(M^{*} > M^{**}\right)} \\ & + \left(M^{*}\right)^2 \left\{M^{*} + n^{-1}\tan^{-1}\left[n(M^{*} - M^{**})\right]\right\} I_{\left(M^{*} < M^{**}\right)}; \end{aligned} \tag{5.40}$$

$$MSE_4^{BOOT}(\theta_i^{EBP}) = \frac{2M^{*}}{1 + \exp\left[2\left(\frac{M^{**}}{M^{*}} - 1\right)\right]}; \tag{5.41}$$

where $I_{(\bullet)}$ is the indicator function which takes the value one when the condition $(\bullet)$ is satisfied, and the value zero otherwise.

When $M^{*} \approx M^{**}$, all the four parametric bootstrap estimators are approximately equal to each other. The first estimator MSE_1^{BOOT} is a straightforward and natural choice. The second and third estimators MSE_2^{BOOT} and MSE_3^{BOOT} were originally considered by Hall and Maiti (2006) for mean-square prediction error calibration. Chatterjee and Lahiri (2008) showed empirically that the last estimator MSE_4^{BOOT} performs marginally better than MSE_2^{BOOT} and MSE_3^{BOOT} and it is reasonably close to the intuitive formula MSE_1^{BOOT}. A desirable property of MSE_4^{BOOT} is that it is always positive. Furthermore, it has the following nice second order accuracy properties (Chatterjee and Lahiri, 2008):

$$E\left[MSE_4^{BOOT}\left(\theta_i^{EBP}\right)\right] = MSE\left(\theta_i^{EBP}\right) + o(d^2 n^{-1}), \tag{5.42}$$

and

$$E\left[MSE_4^{BOOT}\left(\theta_i^{EBP}\right) - MSE\left(\theta_i^{EBP}\right)\right]^2 = O(d^2 n^{-1}), \tag{5.43}$$

where d is the dimension of the hyperparameter $\boldsymbol{\lambda}$ and n is the total sample size. Formula (5.42) shows that the bootstrap estimator is approximately unbiased for the true $MSE(\theta_i^{EBP})$, and formula (5.43) gives the magnitude of the variability of this estimator. Similar properties hold when the other three bootstrap estimators are used.

The implementation of the proposed parametric bootstrap method to estimate $MSE(\theta_i^{EBP})$ is very simple in practice if we know how to estimate a consistent estimator $\tilde{\boldsymbol{\lambda}}$ for one dataset. To estimate M^{*} and M^{**}, all we need to do is to repeat

the first three resampling steps R times, and compute the statistics θ_i^*, $\theta_i^{EBP}(\mathbf{y}_{mn_m}^*, \tilde{\boldsymbol{\lambda}}^*)$, θ_i^{**} and $\theta_i^{EBP}(\mathbf{y}_{mn_m}^{**}, \tilde{\boldsymbol{\lambda}}^{**})$ each time.

Let $\left\{\theta_i^{*(r)},\ \theta_i^{**(r)},\ \theta_i^{EBP}\left(\mathbf{y}_{mn_m}^{*(r)}, \tilde{\boldsymbol{\lambda}}^{*(r)}\right),\ \theta_i^{EBP}\left(\mathbf{y}_{mn_m}^{**(r)}, \tilde{\boldsymbol{\lambda}}^{**(r)}\right),\ \mathrm{r} = 1, ..., R\right\}$ denote all the statistics obtaining from the R resampling processes. If R is large enough, we can estimate M^* and M^{**} using the following formulas:

$$M^* = \frac{1}{R}\sum_{r=1}^{R}\left[\theta_i^{*(r)} - \theta_i^{EBP}\left(\mathbf{y}_{mn_m}^{*(r)},\ \tilde{\boldsymbol{\lambda}}^{*(r)}\right)\right]^2; \tag{5.44}$$

$$M^{**} = \frac{1}{R}\sum_{r=1}^{R}\left[\theta_i^{**(r)} - \theta_i^{EBP}\left(\mathbf{y}_{mn_m}^{**(r)},\ \tilde{\boldsymbol{\lambda}}^{**(r)}\right)\right]^2. \tag{5.45}$$

Once M^* and M^{**} are obtained, the four parametric bootstrap MSE estimators can be computed easily.

5.7 Estimating the MSE for an HB Estimator

In the EBP approach, MSE criteria have been widely used to measure the uncertainty of the EBP of a parameter of interest. However, in the HB approach, the parameters of interest are estimated by the posterior means, and the posterior variances are used as a measure of precision of the estimator, provided they are finite. However, the MSE of EBP and the posterior variance of HB are not comparable. A natural question is: Can we estimate the MSE of the HB estimator? To answer this question, in this section, we study the relationship between $MSE\left(\theta_i^{HB}\right)$ and

$MSE\left(\theta_i^{EBP}\right)$ for the same parameter of interest θ_i based on the same observed data and same model, and then show how to estimate $MSE\left(\theta_i^{HB}\right)$.

When the hyperparameter $\boldsymbol{\lambda}$ is assumed known, both the HB estimator and the EBP estimator of θ_i reduce to the BP of θ_i , that is, $E(\theta_i \mid \mathbf{y}_s, \boldsymbol{\lambda})$, which can be expressed as (4.17) in Chapter 4 or (5.16) in this chapter.

When the hyperparameter $\boldsymbol{\lambda}$ is unknown, the EBP estimator of θ_i is

$$\theta_i^{EBP} = E(\theta_i \mid \mathbf{y}_s, \tilde{\boldsymbol{\lambda}}) = \tilde{\psi}_i(y_i, \tilde{\boldsymbol{\lambda}}), \tag{5.46}$$

where $\tilde{\boldsymbol{\lambda}}$ is a consistent estimator of $\boldsymbol{\lambda}$. As pointed out earlier, ML approach can be used to estimate $\boldsymbol{\lambda}$, and the ML estimator of $\boldsymbol{\lambda}$ is the estimator which maximizes the following likelihood (denoted as L_{EBP}):

$$L_{EBP} = f(\mathbf{y}_s; \boldsymbol{\lambda}) \propto \prod_{i=1}^{m} f(y_i; \boldsymbol{\lambda}) = \prod_{i=1}^{m} \int f(y_i \mid v_i, \boldsymbol{\lambda}) f(v_i \mid \boldsymbol{\lambda}) dv_i. \tag{5.47}$$

Turning to the HB approach, following Kass and Steffey (1989), θ_i^{HB} can be approximated using the first-order Laplace's method as below:

$$\theta_i^{HB} = E(\theta_i \mid \mathbf{y}_s) = E(\theta_i \mid \mathbf{y}_s, \hat{\boldsymbol{\lambda}}) \left[1 + O(m^{-1})\right], \tag{5.48}$$

where $\hat{\boldsymbol{\lambda}}$ is the posterior mode. According to the definition, $\hat{\boldsymbol{\lambda}}$ is the estimator which maximizes the following function (denoted as L_{HB}):

$$L_{HB} = f(\mathbf{y}_s; \boldsymbol{\lambda}) \propto \prod_{i=1}^{m} f(y_i; \boldsymbol{\lambda}) = \prod_{i=1}^{m} \int f(y_i \mid v_i, \boldsymbol{\lambda}) f(v_i \mid \boldsymbol{\lambda}) \pi(\boldsymbol{\lambda}) dv_i, \tag{5.49}$$

where $\pi(\boldsymbol{\lambda})$ is the subjective prior distribution of $\boldsymbol{\lambda}$.

Assume $\pi(\boldsymbol{\lambda}) \propto c$, where c is some constant, then the two functions defined by (5.47) and (5.49) are proportional to each other, that is $L_{HB} \propto L_{EBP}$. As a result, the MLE estimator $\tilde{\boldsymbol{\lambda}}$ for the EBP approach is equal to the posterior mode $\hat{\boldsymbol{\lambda}}$ for the HB approach, that is, $\hat{\boldsymbol{\lambda}} = \tilde{\boldsymbol{\lambda}}$. Combining this result with (5.46) and (5.48), we obtain the following result:

$$\theta_i^{HB} = \theta_i^{EBP}\left[1 + O(m^{-1})\right]. \tag{5.50}$$

Thus,

$$\begin{aligned} MSE\left(\theta_i^{HB}\right) &= MSE\left\{\theta_i^{EBP}\left[1 + O(m^{-1})\right]\right\} \\ &= MSE\left(\theta_i^{EBP}\right)\left[1 + O(m^{-2})\right]. \end{aligned} \tag{5.51}$$

Therefore, when $\pi(\lambda) \propto c$ is assumed for the HB, $MSE\left(\theta_i^{HB}\right)$ can be approximated by $MSE\left(\theta_i^{EBP}\right)$ within order $O_p(m^{-2})$, where ML approach is used to estimate $\boldsymbol{\lambda}$ for the EBP.

5.8 Concluding Remarks

We have studied how to make inferences using the EBP approach for our proposed *Bernoulli-Logit-EP* model. Both the BP and the EBP for the random effect v_i and the mixed effect θ_i have been presented. We have also developed a methodology for estimating the MSE of the EBP using a Taylor series linearization approach and a double parametric bootstrap approach. Finally we have shown how to estimate the MSE of the HB estimation under certain conditions. We have not attempted to provide a rigorous proof of the exact order the remainder terms. The

advantages of the proposed double parametric bootstrap approach include simplicity of implementation and guarantee that the estimator will be positive.

Chapter 6: Adaptive Hierarchical Bayes Estimation of Small-Area Proportions under Two-stage Sampling

6.1 Introduction

In the previous chapters, we have considered various methodologies for estimating small area proportions when the samples are drawn from a single stage design. In this chapter, we extend the ideas by estimating the small-area proportions when samples are drawn using a two-stage sample design.

This chapter proposes a generalized linear mixed model that is suitable for binary data collected from a two-stage sample design. Like the previous chapters, in order to allow for kurtosis in the random effects, we relax the normality assumption for the random effects to allow for a class of distributions. The model and notation are presented in Section 6.2. In Section 6.3, we illustrate some Bayesian inferences based on the proposed model. In Section 6.4, we conduct a data analysis using the proposed model based on samples drawn from a real finite population. The chapter finishes with some concluding remarks in Section 6.5.

6.2 Notation and Model

The small areas of interest are geographic areas that contain one or more primary sampling units (PSUs). Suppose that the i th small area contains C_i PSUs and the j th PSU in the i th area contains N_{ij} secondary units (elements). Let y_{ijk} be the binary characteristic of interest associated with unit k in PSU j of area i

($k = 1,...,N_{ij};\ j = 1,...,C_i;\ i = 1,...,m$). A column vector of p known covariates, $\mathbf{x}_{ij} = (x_{ij1},\ ...,\ x_{ijp})'$, is assumed to be the same for each unit k in PSU j of area *i*.

Under the above population structure, a common practice is to employ two-stage sample design in each small area: Assume c_i PSUs are selected from area i; if the *j* th PSU is sampled, n_{ij} elements are then selected from that PSU. If we know the small areas of interest at the design stage, it is possible to make them into strata, in which case the number of sampled PSUs c_i at the first stage is fixed and positive. Similar designs were considered by Scott and Smith (1969), Ghosh and Lahiri (1988), and Stukel and Rao (1999).

However, in many cases, the small areas of interest are not identified at the design stage, and therefore they may not be design strata. Also, there may be too many small areas to make each of them into strata. As a result, the number of sampled PSUs falling into a small area is random and can be zero (e.g., see Mohadjer et al., 2007). In this research, we focus on this type of situation.

For simplicity, we consider a two-stage design where a PSUs are selected with probabilities proportional to size measures (PPS) from the PSU frame at the first stage, and then a set of elements of size b is selected with equal probability at the second stage within each sampled PSU. With this design, all the sampled elements are selected with equal overall probability (EPSEM). We still let c_i denote the number of sampled PSUs falling into area i. Then $0 \le c_i \le C_i$ and c_i is random, $i = 1,...,m$.

Inferences about the finite population small-area proportions P_i are to be considered, where $P_i = \Sigma_{j=1}^{C_i} \Sigma_{k=1}^{N_{ij}} y_{ijk} \Big/ \Sigma_{j=1}^{C_i} N_{ij}$, $i = 1,...,m$. Bayesian inference for P_i requires assumptions about the distribution of y_{ijk} for $i = 1,...,m;\ j = 1,...,C_i;\ k = 1,...,N_{ij}$ and a prior distribution for the parameters of the sampling distribution of $\left\{ y_{ijk} \right\}$. Independently for all $(i,\ j,\ k)$, it is reasonable to assume that

$$y_{ijk} \mid \theta_{ij} \overset{ind}{\sim} Bernoulli(\theta_{ij}), \quad i = 1,...,m;\ j = 1,...,C_i;\ k = 1,...,N_{ij}. \tag{6.1}$$

Logistic regression with mixed effects is often assumed for the prior distribution of $\left\{ \theta_{ij} : i = 1,...,m;\ j = 1,...,C_i \right\}$. That is,

$$\text{logit}(\theta_{ij}) = \mathbf{x}'_{ij}\boldsymbol{\beta} + v_i + u_{ij}, \quad i = 1,...,m;\ j = 1,...,C_i, \tag{6.2}$$

whereas, independently for all $(i,\ j)$,

$$u_{ij} \overset{iid}{\sim} \text{N}(0,\ \sigma_u^2), \quad i = 1,...,m;\ j = 1,...,C_i, \tag{6.3}$$

and independently for all i,

$$v_i \overset{iid}{\sim} \text{N}(0,\ \sigma_v^2), \quad i = 1,...,m. \tag{6.4}$$

The effect u_{ij} accounts for the sampling design and v_i accounts for the random area effect. We call the model defined by (6.1)~(6.4) the *Bernoulli-Logit-Normal* model.

As discussed in Chapter 3, the distributions of the random effects for some data may depart from normality. Instead of the normality assumption, we assume a

class of distributions – the exponential power distributions – for the random effects v_i and u_{ij}. That is, the following models are assumed for v_i and u_{ij}:

$$v_i \overset{iid}{\sim} EP(0, \sigma_v, \varphi_v), \quad i = 1, ..., m, \tag{6.5}$$

and

$$u_{ij} \overset{iid}{\sim} EP(0, \sigma_u, \varphi_u), \quad i = 1, ..., m; \; j = 1, ..., C_i, \tag{6.6}$$

where the hyperparameters σ_v, σ_u, φ_v and φ_u are assumed unknown. We call the model defined by (6.1)~(6.2) & (6.5)~(6.6) the *Bernoulli-Logit-EP* model. Again, the strength of this model is that it uses a class of probability distributions instead of a specific one, and the underlying model will be chosen adaptively by the data.

6.3 Bayesian Inference

Let s_i denote the sample of PSUs and s_{ij} denote the sample of elements, $j = 1, ..., c_i$, $i = 1, ..., m$. Let $\mathbf{y}_s = (y_{11}, ..., y_{1c_1}; ...; y_{m1}, ..., y_{mc_m})'$ denote the vector of the PSU level sample totals, where $y_{ij} = \sum_{k=1}^{n_{ij}} y_{ijk}$ and n_{ij} is the sample size for PSU j in area i.

The Bayes estimator of P_i is the mean of the posterior distribution of P_i. First, consider the areas that contain at least one sampled PSU, i.e., $c_i > 0$. We can rewrite P_i as

$$P_i = \frac{1}{N_i}\left(\Sigma_{j \in s_i} \Sigma_{k \in s_{ij}} y_{ijk} + \Sigma_{j \in s_i} \Sigma_{k \in s_{ij}^c} y_{ijk} + \Sigma_{j \in s_i^c} \Sigma_{k=1}^{N_{ij}} y_{ijk} \right),$$

where s_{ij}^c is the set of non-sampled elements in the j th sampled PSU, s_i^c is the set of non-sampled PSUs, and $N_i = \Sigma_{j=1}^{C_i} N_{ij}$ is the area level population total.

From the assumed Bernoulli component of the model, $E\left(y_{ijk} \mid \theta_{ij}\right) = \theta_{ij}$ and $V\left(y_{ijk} \mid \theta_{ij}\right) = \theta_{ij}(1-\theta_{ij})$. Therefore,

$$\begin{aligned} E\left(P_i \mid \theta_{ij}, \mathbf{y}_s\right) &= \frac{1}{N_i}\left[\Sigma_{j\in s_i}\Sigma_{k\in s_{ij}} y_{ijk} + \Sigma_{j\in s_i}\Sigma_{k\in s_{ij}^c} E\left(y_{ijk} \mid \theta_{ij}, \mathbf{y}_s\right)\right] \\ &\quad + \frac{1}{N_i}\Sigma_{j\in s_i^c}\Sigma_{k=1}^{N_{ij}} E\left(y_{ijk} \mid \theta_{ij}, \mathbf{y}_s\right) \qquad (6.7) \\ &= \frac{1}{N_i}\left[\Sigma_{j\in s_i}\Sigma_{k\in s_{ij}} y_{ijk} + \Sigma_{j\in s_i}(N_{ij}-n_{ij})\theta_{ij} + \Sigma_{j\in s_i^c} N_{ij}\theta_{ij}\right], \end{aligned}$$

and

$$\begin{aligned} V\left(P_i \mid \theta_{ij}, \mathbf{y}_s\right) &= \frac{1}{N_i^2}\left\{V\left[\left(\Sigma_{j\in s_i}\Sigma_{k\in s_{ij}^c} y_{ijk}\right)\middle|\theta_{ij}, \mathbf{y}_s\right] + V\left[\left(\Sigma_{j\in s_i^c}\Sigma_{k=1}^{N_{ij}} y_{ijk}\right)\middle|\theta_{ij}, \mathbf{y}_s\right]\right\} \\ &= \frac{1}{N_i^2}\left[\Sigma_{j\in s_i}\Sigma_{k\in s_{ij}^c} V(y_{ijk} \mid \theta_{ij}, y_s) + \Sigma_{j\in s_i^c}\Sigma_{k=1}^{N_{ij}} V\left(y_{ijk} \mid \theta_{ij}, y_s\right)\right] \qquad (6.8) \\ &= \frac{1}{N_i^2}\left[\Sigma_{j\in s_i}(N_{ij}-n_{ij})\theta_{ij}(1-\theta_{ij}) + \Sigma_{j\in s_i^c} N_{ij}\theta_{ij}(1-\theta_{ij})\right]. \end{aligned}$$

Hence, the posterior mean of P_i is:

$$\begin{aligned} E(P_i \mid \mathbf{y}_s) &= E\left[E(P_i \mid \theta_{ij}, \mathbf{y}_s) \mid \mathbf{y}_s\right] \\ &= \frac{1}{N_i} E\left\{\left[\Sigma_{j\in s_i}\Sigma_{k\in s_{ij}} y_{ijk} + \Sigma_{j\in s_i}(N_{ij}-n_{ij})\theta_{ij} + \Sigma_{j\in s_i^c} N_{ij}\theta_{ij}\right]\middle|\mathbf{y}_s\right\} \qquad (6.9) \\ &= \frac{1}{N_i}\left[\Sigma_{j\in s_i}\Sigma_{k\in s_{ij}} y_{ijk} + \Sigma_{j\in s_i}(N_{ij}-n_{ij})E(\theta_{ij} \mid \mathbf{y}_s) + \Sigma_{j\in s_i^c} N_{ij}E(\theta_{ij} \mid \mathbf{y}_s)\right], \end{aligned}$$

and the posterior variance of P_i is:

$$
\begin{aligned}
V(P_i \mid \mathbf{y}_s) &= E\left[V\left(P_i \mid \mathbf{y}_s, \theta_{ij}\right) \mid \mathbf{y}_s\right] + V\left[E\left(P_i \mid \mathbf{y}_s, \theta_{ij}\right) \mid \mathbf{y}_s\right] \\
&= \frac{1}{N_i^2} E\left\{\left[\Sigma_{j\in s_i}(N_{ij}-n_{ij})\theta_{ij}(1-\theta_{ij}) + \Sigma_{j\in s_i^c} N_{ij}\theta_{ij}(1-\theta_{ij})\right] \middle| \mathbf{y}_s\right\} \\
&\quad + \frac{1}{N_i^2} V\left\{\left[\Sigma_{j\in s_i}\Sigma_{k\in s_{ij}} y_{ijk} + \Sigma_{j\in s_i}(N_{ij}-n_{ij})\theta_{ij} + \Sigma_{j\in s_i^c} N_{ij}\theta_{ij}\right] \middle| \mathbf{y}_s\right\} \qquad (6.10) \\
&= \frac{1}{N_i^2}\left\{\Sigma_{j\in s_i}(N_{ij}-n_{ij}) E\left[\theta_{ij}(1-\theta_{ij}) \mid \mathbf{y}_s\right] + \Sigma_{j\in s_i^c} N_{ij} E\left[\theta_{ij}(1-\theta_{ij}) \mid \mathbf{y}_s\right]\right\} \\
&\quad + \frac{1}{N_i^2} V\left\{\left[\Sigma_{j\in s_i}(N_{ij}-n_{ij})\theta_{ij} + \Sigma_{j\in s_i^c} N_{ij}\theta_{ij}\right] \mid \mathbf{y}_s\right\},
\end{aligned}
$$

where $\theta_{ij} = \dfrac{\exp\left(\mathbf{x}_{ij}'\boldsymbol{\beta} + v_i + u_{ij}\right)}{1+\exp\left(\mathbf{x}_{ij}'\boldsymbol{\beta} + v_i + u_{ij}\right)}$. Note that the posterior mean $E(\theta_{ij} \mid \mathbf{y}_s)$ for $j \in s_i$ is different from $E(\theta_{ij} \mid \mathbf{y}_s)$ for $j \in s_i^c$. It is difficult to express the posterior means of θ_{ij} in a closed-form because complicated integrations are involved.

Now, consider the areas that do not contain any sampled PSU, i.e., $c_i = 0,$ for some area i. For these areas, we can rewrite P_i as $P_i = \dfrac{1}{N_i}\left(\Sigma_{j=1}^{C_i}\Sigma_{k=1}^{N_{ij}} y_{ijk}\right)$. Then the posterior mean and posterior variance of P_i become a special case of (6.9) and (6.10) respectively. That is:

$$
E(P_i \mid \mathbf{y}_s) = \frac{1}{N_i}\left[\Sigma_{j=1}^{C_i} N_{ij} E(\theta_{ij} \mid \mathbf{y}_s)\right], \qquad (6.11)
$$

$$
V(P_i \mid \mathbf{y}_s) = \frac{1}{N_i^2}\left\{\Sigma_{j=1}^{C_i} N_{ij} E\left[\theta_{ij}(1-\theta_{ij}) \middle| \mathbf{y}_s\right] + V\left[\left(\Sigma_{j=1}^{C_i} N_{ij}\theta_{ij}\right) \middle| \mathbf{y}_s\right]\right\}. \qquad (6.12)
$$

Let $\boldsymbol{\theta} = (\theta_{11},...,\theta_{1c_1};...;\theta_{m1},...,\theta_{mc_m})'$ and $\mathbf{v} = (v_1,...,v_m)'$. Let $\boldsymbol{\lambda}_{Norm} = \left(\boldsymbol{\beta}, \sigma_v, \sigma_u\right)'$ and $\boldsymbol{\lambda}_{EP} = \left(\boldsymbol{\beta}, \sigma_v, \sigma_u, \varphi_v, \varphi_u\right)'$ denote the hyperparameters for

the *Bernoulli-Logit-Normal* model and the *Bernoulli-Logit-EP* model respectively. Reasonable prior assumptions will be applied to all the hyperparameters $\boldsymbol{\lambda}_{Norm}$ and $\boldsymbol{\lambda}_{EP}$.

The joint posterior distribution of all the model parameters based on the *Bernoulli-Logit-Normal* model is:

$$\begin{aligned} f(\boldsymbol{\theta},\mathbf{v},\boldsymbol{\lambda}_{Norm}\,|\,\mathbf{y}_s) &\propto f(\mathbf{y}_s\,|\,\boldsymbol{\theta},\mathbf{v},\boldsymbol{\lambda}_{Norm})f(\boldsymbol{\theta}\,|\,\mathbf{v},\boldsymbol{\lambda}_{Norm})f(\mathbf{v}\,|\,\sigma_v)f(\boldsymbol{\lambda}_{Norm}) \\ &\propto f(\lambda_{Norm})\prod_{i=1}^{m}\left\{\prod_{j=1}^{c_i}\theta_{ij}^{y_{ij}-1}(1-\theta_{ij})^{n_{ij}-y_{ij}-1}\frac{1}{\sigma_u}\exp\left[-\frac{\left(\text{logit}(\theta_{ij})-\mathbf{x}_{ij}'\boldsymbol{\beta}-v_i\right)^2}{2\sigma_u^2}\right]\right\} \\ &\quad\times\left[\frac{1}{\sigma_v}\exp\left(-\frac{v_i^2}{2\sigma_v^2}\right)\right]. \end{aligned}$$

(6.13)

The joint posterior distribution of all the model parameters based on the *Bernoulli-Logit-EP* model is:

$$\begin{aligned} f\left(\boldsymbol{\theta},\mathbf{v},\boldsymbol{\lambda}_{EP}\middle|\mathbf{y}_s\right) &\propto f(\mathbf{y}_s\,|\,\boldsymbol{\theta},\mathbf{v},\boldsymbol{\lambda}_{EP})f(\boldsymbol{\theta}\,|\,\mathbf{v},\boldsymbol{\lambda}_{EP})f(\mathbf{v}\,|\,\sigma_v)f(\boldsymbol{\lambda}_{EP}) \\ &\propto f(\boldsymbol{\lambda}_{EP})\prod_{i=1}^{m}\left[\prod_{j=1}^{c_i}\theta_{ij}^{y_{ij}-1}(1-\theta_{ij})^{n_{ij}-y_{ij}-1}\frac{c_{1u}}{\sigma_u}\exp\left(-\frac{c_{0u}^{1/(2\varphi_u)}\left|\text{logit}(\theta_{ij})-\mathbf{x}_{ij}'\boldsymbol{\beta}-v_i\right|^{1/\varphi_u}}{\sigma_u^{1/\varphi_u}}\right)\right] \\ &\quad\times\left[\frac{c_{1v}}{\sigma_v}\exp\left(-\frac{c_{0v}^{1/(2\varphi_v)}|v_i|^{1/\varphi_v}}{\sigma_v^{1/\varphi_v}}\right)\right], \end{aligned}$$

(6.14)

where $c_{0u}=\Gamma(3\varphi_u)/\Gamma(\varphi_u)$, $c_{1u}=\sqrt{c_{0u}}\,/\left[2\varphi_u\Gamma(\varphi_u)\right]$, $c_{0v}=\Gamma(3\varphi_v)/\Gamma(\varphi_v)$, and $c_{1v}=\sqrt{c_{0v}}\,/\left[2\varphi_v\Gamma(\varphi_v)\right]$.

Neither of the joint posterior distributions defined by (6.13) and (6.14) can be expressed in a closed form, and therefore approximations are needed. However, the

joint posterior distributions can be simulated using MCMC methods, which can be implemented using the Gibbs sampler or the Metropolis-Hastings algorithm. We will implement the HB models using the MCMC technique in this chapter.

In order to estimate $E(P_i \mid \mathbf{y}_s)$ and $V(P_i \mid \mathbf{y}_s)$ as defined by (6.9) and (6.10), one can follow the approach used by Malec et al. (1997). Let $\boldsymbol{\lambda}$ denote the hyperparameters, where $\boldsymbol{\lambda} = \boldsymbol{\lambda}_{Norm}$ for the *Bernoulli-Logit-Normal* model, and $\boldsymbol{\lambda} = \boldsymbol{\lambda}_{EP}$ for the *Bernoulli-Logit-EP* model. Following Malec et al. (1997), we can first use the Metropolis-Hastings algorithm within the Gibbs sampler to generate R sets of parameters

$$\left\{\left[\left(\theta_{ij}^{(r)} : j \in s_i, i = 1,...,m\right), \left(\theta_{ij(RD)}^{(r)} : j \in s_i^c, i = 1,...,m\right), \mathbf{v}^{(r)}, \boldsymbol{\lambda}^{(r)}\right] : r = 1,...,R\right\}$$ from

their full conditional distributions (see the Appendix for the full conditional distributions of all the model parameters for both models), where $\theta_{ij}^{(r)}$ and $\theta_{ij(RD)}^{(r)}$ denote the MCMC values for the sampled PSUs and nonsampled PSUs in area i respectively. Note that the full conditional distributions from which $\theta_{ij}^{(r)}$ and $\theta_{ij(RD)}^{(r)}$ were drawn are different. Then we can use the R sets, $\left\{\theta_{ij}^{(r)};\ j \in s_i,\ i = 1,...,m,\ r = 1,...,R\right\}$ and $\left\{\theta_{ij(RD)}^{(r)} : j \in s_i^c,\ i = 1,...,m,\ r = 1,...,R\right\}$, to obtain estimates of $E\left(P_i \mid \mathbf{y}_s\right)$ and $V\left(P_i \mid \mathbf{y}_s\right)$ as follows:

$$\begin{aligned} P_i^{HB} &= \hat{E}\left(P_i \mid \mathbf{y}_s\right) \\ &= \frac{1}{N_i}\left\{\Sigma_{j \in s_i}\Sigma_{k \in s_{ij}} y_{ijk} + R^{-1}\sum_{r=1}^{R}\left[\Sigma_{j \in s_i}(N_{ij} - n_{ij})\theta_{ij}^{(r)} + \Sigma_{j \in s_i^c} N_{ij}\theta_{ij(RD)}^{(r)}\right]\right\}; \end{aligned} \tag{6.15}$$

$$
\begin{aligned}
V(P_i^{HB}) &= \hat{V}\left(P_i \mid \mathbf{y}_s\right) \\
&= \frac{1}{N_i^2 R}\sum_{r=1}^{R}\left\{\Sigma_{j\in s_i}(N_{ij}-n_{ij})\theta_{ij}^{(r)}[1-\theta_{ij}^{(r)}]+\Sigma_{j\in s_i^c}N_{ij}\theta_{ij(RD)}^{(r)}[1-\theta_{ij(RD)}^{(r)}]\right\} \\
&+ \frac{1}{N_i^2 R}\sum_{r=1}^{R}\left[\Sigma_{j\in s_i}(N_{ij}-n_{ij})\theta_{ij}^{(r)}+\Sigma_{j\in s_i^c}N_{ij}\theta_{ij(RD)}^{(r)}\right]^2 \\
&- \frac{1}{N_i^2 R}\left\{\sum_{r=1}^{R}\left[\Sigma_{j\in s_i}(N_{ij}-n_{ij})\theta_{ij}^{(r)}+\Sigma_{j\in s_i^c}N_{ij}\theta_{ij(RD)}^{(r)}\right]\right\}^2 .
\end{aligned}
\tag{6.16}
$$

One disadvantage of the above approach is that it produces only posterior means and posterior variances. It is, for example, not possible to compute credible intervals. To avoid this disadvantage, we propose a fully Bayesian approach for the small area finite population proportion P_i.

A fully Bayesian approach is to generate MCMC values $\left(P_i^{(r)} : r = 1,..., R\right)$ for P_i as follows:

$$
P_i^{(r)} = \frac{1}{N_i}\left(\Sigma_{j\in s_i}\Sigma_{k\in s_{ij}}y_{ijk} + \Sigma_{j\in s_i}y_{ij}^{(r)} + \Sigma_{j\in s_i^c}y_{ij(RD)}^{(r)}\right), \quad r = 1,..., R, \tag{6.17}
$$

where $y_{ij}^{(r)} \sim Bin\left(N_{ij}-n_{ij},\ \theta_{ij}^{(r)}\right)$ for $j \in s_i$ and $y_{ij(RD)}^{(r)} \sim Bin\left(N_{ij},\ \theta_{ij(RD)}^{(r)}\right)$ for $j \in s_i^{(c)}$. The values

$$
\left\{\left[\left(\theta_{ij}^{(r)}, j\in s_i, i = 1,..., m\right), \left(\theta_{ij(RD)}^{(r)}, j\in s_i^c, i = 1,..., m\right), \mathbf{v}^{(r)}, \boldsymbol{\lambda}^{(r)}\right] : r = 1,..., R\right\}
$$

are obtained from their full conditional distributions (see the Appendix) using the Metropolis-Hastings algorithm within the Gibbs sampler algorithm. Note that for the areas without any sample, $P_i^{(r)} = \frac{1}{N_i}\left[\Sigma_{j=1}^{C_i}y_{ij(RD)}^{(r)}\right]$.

The posterior mean and variance for P_i can be estimated as follows:

$$P_i^{HB} = R^{-1}\sum_{r=1}^{R} P_i^{(r)}; \tag{6.18}$$

and

$$V(P_i^{HB}) = R^{-1}\sum_{r=1}^{R}\left(P_i^{(r)} - P_i^{HB}\right)^2. \tag{6.19}$$

Note that the posterior mean and posterior variance of P_i estimated from formulas (6.18) and (6.19) should be very close to those estimated using (6.15) and (6.16) respectively. Using the fully Bayesian method, we can also compute the credible intervals for P_i using the MCMC values $\left\{P_i^{(r)} : r = 1,..., R\right\}$.

6.4 Data Analysis

6.4.1 The Study Population and the Sample Design

As in earlier chapters, we considered the 2002 Natality public-use data file as our study population. The file included all births occurring within the United States in 2002. The objective was to draw a set of samples from the finite population using a two-stage sample design, where single counties were treated as PSUs. Since for confidentiality reason, only counties of 100,000 or more population based on the 1990 Census could be identified in the 2002 Natality data, we restricted our study population to those counties. As a result, our finite study population comprised of 3,270,509 live-birth records in the counties of 100,000 or more population based on the 1990 Census. The final population included live-birth records in 454 counties within 49 states plus the District of Columbia (Wyoming was excluded). The

parameter of interest was the state level low birth weight rate P_i, $i = 1,...,50$, where low birth weight was defined as birth weight less than 2,500 grams. The values of P_i varied from 4.9 percent to 15.1 percent across the states in this finite population.

To mimic a real survey design, at the first sampling stage, our objective was to select 80 PSUs with PPS from the PSU frame. Let N_j denote the number of live-birth records in the population of the j th PSU. Treating N_j as the size measure for the PPS selection, the probability for the j th PSU to be selected would be $prob_j = cN_j \Big/ \Sigma_{j=1}^{454} N_j$, where $c = 80$ is the sample size for the PSU selection, $c = 1,...,454$. The probabilities $prob_j$ for eight PSUs exceeded 1. We therefore selected those eight PSUs with certainty. SAS PROC SURVEYSELECT (METHOD=PPS) was then used to select the remaining 72 PSUs from the remaining 446 PSUs on the frame without replacement. The procedure of "PPS sampling without replacement" in SAS uses the Hanurav-Vijayan algorithm for PPS selection without replacement (Hanurav, 1967; Vijayan, 1968). According to the Hanurav-Vijayan algorithm, PROC SURVEYSELECT ordered the PSUs in ascending order by size measure before selecting the units. For details on the selection procedure, we refer to the SAS support website at

http://support.sas.com/onlinedoc/913/docMainpage.jsp.

The first stage sampling probabilities for the sampled PSUs were:

$$prob_j = \begin{cases} 1, & \text{if PSU } j \text{ is selected with certainty} \\ 72N_j \Big/ \Sigma_{j=1}^{446} N_j, & \text{if PSU } j \text{ is selected without certainty} \end{cases}$$

At the second sampling stage, 30 live-birth records were selected using EPSEM within each of the sampled noncertainty PSUs. The second stage sampling probability for the elements within the noncertainty PSU j is therefore $prob_{\beta} = 30 / N_j$. As a result, the unconditional probability for the elements selected in the noncertainty PSUs was: $prob_u = \frac{72 \times 30}{\sum_{j=1}^{446} N_j}$. For the certainty PSUs, the first stage sampling probability was truncated from some numbers larger than 1 to 1, therefore the second stage sampling probability should be increased to retain the original conditional probability. In order to retain an EPSEM design, the sampling rate within each certainty PSU α' was set to $prob_u$. SAS PROC SURVEYSELECT (METHOD=SRS) was used to select the second stage samples. The resulting unconditional probability for elements selected in the certainty PSUs was equal to $prob_u$.

The final samples were from 80 sampled PSUs within 32 states. The number of selected PSUs within each of the 32 states varied from 1 to 12. Among those states, 15 states contained at least two sampled PSUs including the certainties, and the other 17 states contained only one sampled PSU. The state level sample sizes n_i varied from 30 to 476, and the PSU level sample sizes n_{ij} varied from 30 to 129. The other 18 states did not contain any samples.

Our objective was to estimate the state level low birthweight rate P_i, $i = 1, ..., 50$. Based on the samples selected using the design described in section

6.4.1, we can only compute the direct point estimate of P_i for the 32 states with samples as follows:

$$p_{iw} = \frac{\sum_{j=1}^{c_i} \sum_{k=1}^{n_{ij}} y_{ijk}}{\sum_{j=1}^{c_i} n_{ij}}, \quad i = 1, ..., 32. \tag{6.20}$$

Note that we do not need the survey weights in the computation of p_{iw} because the two-stage design employed is EPSEM. For the states containing at least one non-certainty PSU, the denominator of (6.20) is random because c_i is random, therefore p_{iw} defined by (6.20) are actually ratio estimators. For the states containing only certainty PSUs, the denominator of (6.20) is fixed, therefore p_{iw} are sample means. The values of p_{iw} varied from zero percent to 23.3 percent. Two of the 32 states had zero point estimates. Variance estimation for p_{iw} using design-based approach was not conducted due to too few selected PSUs within most of the states.

To obtain estimates of P_i for the states that had no samples, synthetic estimators or model based/assisted estimators could be considered. To improve the direct estimates for the 32 states that had samples and to predict the estimates for the 18 states that had no samples, HB approach described in Section 6.3 was implemented.

6.4.2 HB Modeling Implementation

Based on the proposed *Bernoulli-Logit-EP* model, we obtained the HB estimates for the finite population proportion P_i using the fully Bayesian approach.

For comparison purpose, we also obtained the HB estimates using the *Bernoulli-Logit-Normal* model. We incorporated two covariates at the PSU level (percentage of births with mothers of age less than 15 and percentage of births that are the first child in the family) in the HB models. These two covariates are the same as the ones used in Chapter 3, except that they are computed at the PSU level instead of the state level. In addition, the following prior distributions were assumed for the hyperparameters:

1) Flat prior for $\boldsymbol{\beta}$: $f(\boldsymbol{\beta}) \propto 1$;
2) Uniform prior for the variance components: $\sigma_v \sim Unif(0,\ L)$ and $\sigma_u \sim Unif(0,\ L)$, where L is a large positive number;
3) Uniform prior for the kurtosis parameters: $\varphi_v \sim Unif(0,\ 1)$ and $\varphi_u \sim Unif(0,\ 1)$.

The HB models were implemented using WinBUGS. For each model, three independent chains were used. For each chain, burn-ins of 20,000 samples were produced, with 20,000 samples after burn-in. The samples after burn-in were thinned by a factor of two to reduce auto-correlation of the MCMC results. The resultant 30,000 MCMC samples after burn-in were then used to compute the posterior mean, variance, and percentiles for each HB model. The potential scale reduction factor $\hat{R}$ was used as the primary measure for convergence (see Gelman and Rubin, 1992).

6.4.3 Comparison of Different Estimation Methods

This subsection compares the results from different estimation methods for the states with and without samples separately. Figures 6-1 to 6-3 present the different

point estimates, associated square root of the posterior variance, and the 95% credible intervals from the *Bernoulli-Logit-EP* model for the 32 states with sampled births, while Figures 6-4 to 6-6 present the same statistics for the 18 states that had no sampled births. In Figures 6-1 to 6-3, the data are sorted in the increasing order of the state level sample size first, and then by the true P_i. In Figures 6-4 to 6-5, the data are sorted by the increasing order of the true P_i.

In Figure 6-1a, we plot three point estimates of the state level low birthweight rates along with the true P_i (true.P) for the 32 states that contained sample births. The three point estimates include: the direct estimates (direct.P), the HB estimates using the *Bernoulli-Logit-Normal* model (HBNorm), and the HB estimates using the *Bernoulli-Logit-EP* model (HBEP). The first 17 states (in the first panel) on the graph had sample sizes $n_i = 30$, $i = 1,...,17$. The next five states (in the second panel) on the graph had sample sizes $n_i = 60$, $i = 18,...,22$. The remaining 10 states (in the last panel) had sample sizes n_i varying from 79 to 476, $i = 23,...,32$. We also computed the residuals of each of the point estimates which are defined as the differences between the estimates and the corresponding true values, i.e., $res_i = estimate(P_i) - P_i$. Figure 6-1b presents the plots of these residuals. The graphs clearly show that the direct estimate performs much worse than the two HB estimates for the states with small sample sizes. The direct estimates do not perform well for many of the states in the first two panels of the graph, where the state sample sizes are relatively small ($n_i = 30$ or 60). The performance improves for the states in the last panel of the graph, where the state sample sizes are all not smaller than 79. It is hard to distinguish

any differences between the two HB estimates from this graph. Further analysis using measure of absolute relative deviation was conducted. We present the results at the later part of this subsection.

Figure 6-2 exhibits the posterior standard errors associated with the two HB estimates. Again, the plot shows that the two HB estimators perform very similarly.

Figure 6-3 displays the 95% credible intervals based on the *Bernoulli-Logit-EP* model. The widths of the credible intervals are smallest in the last panel of the graph. Several states in the middle of the graph (e.g., states 13, 14, 15, 17 and 19) have largest credible interval widths. Many of the credible intervals seem right skewed with respect to the location of the true P_i. This is partly due to the fact of small sample sizes. All the 32 credible intervals cover the true P_i. Note that the credible interval for one state (state 31st) in the last panel covers the true value at the left bound. The credible intervals based on the *Bernoulli-Logit-Normal* model look similar to those displayed in Figure 6-3; we do not display them here.

We did not detect any evidence from Figures 6-1 to 6-3 which could explain the relationship between the different estimates and the true P_i.

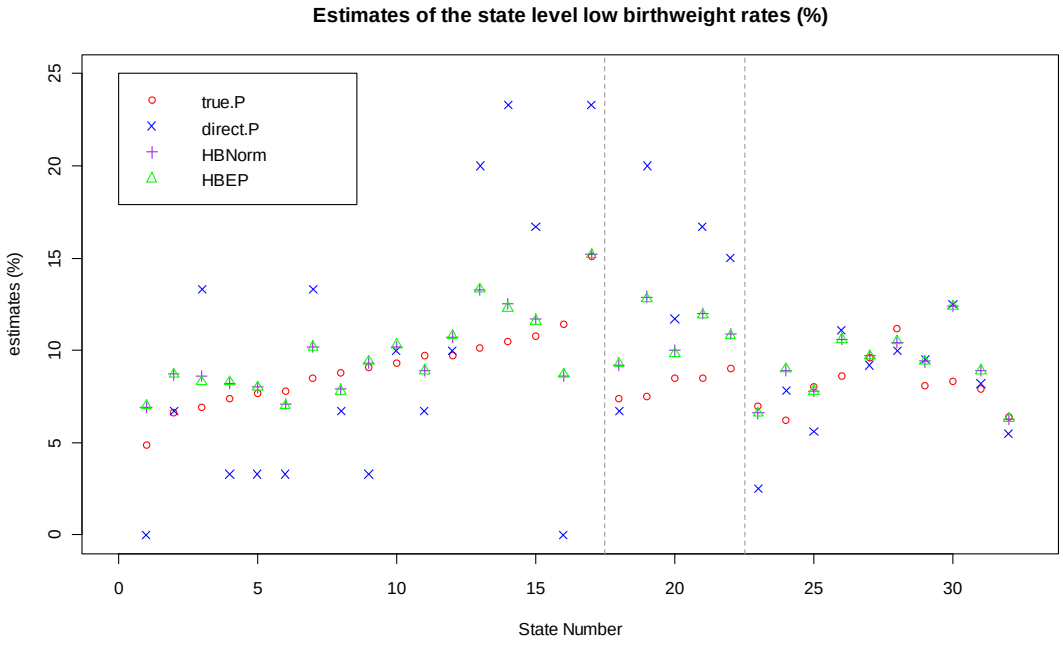

Figure 6-1a: Comparison of different point estimates for the low birthweight rates (in percentages) for the 32 states with sampled births, where states are sorted by the sample sizes and true proportions P_i.

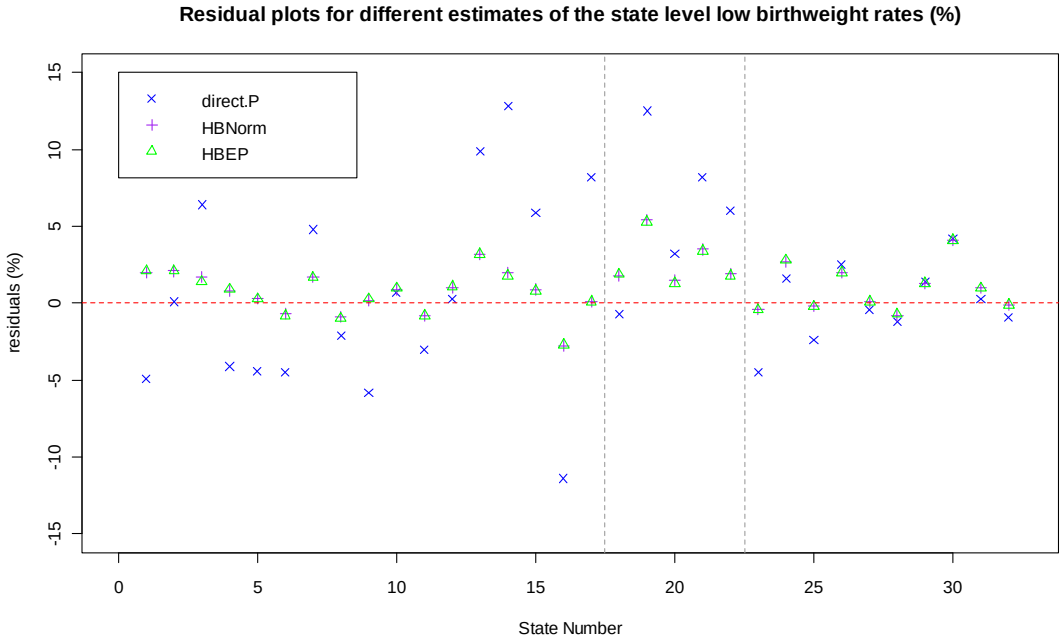

Figure 6-1b: Residual plots of different point estimates for low birthweight rates (in percentags) for the 32 states with sampled births, where the residuals were defined by $estimate(P_i) - P_i$, and states were sorted by the sample sizes and true proportions P_i.

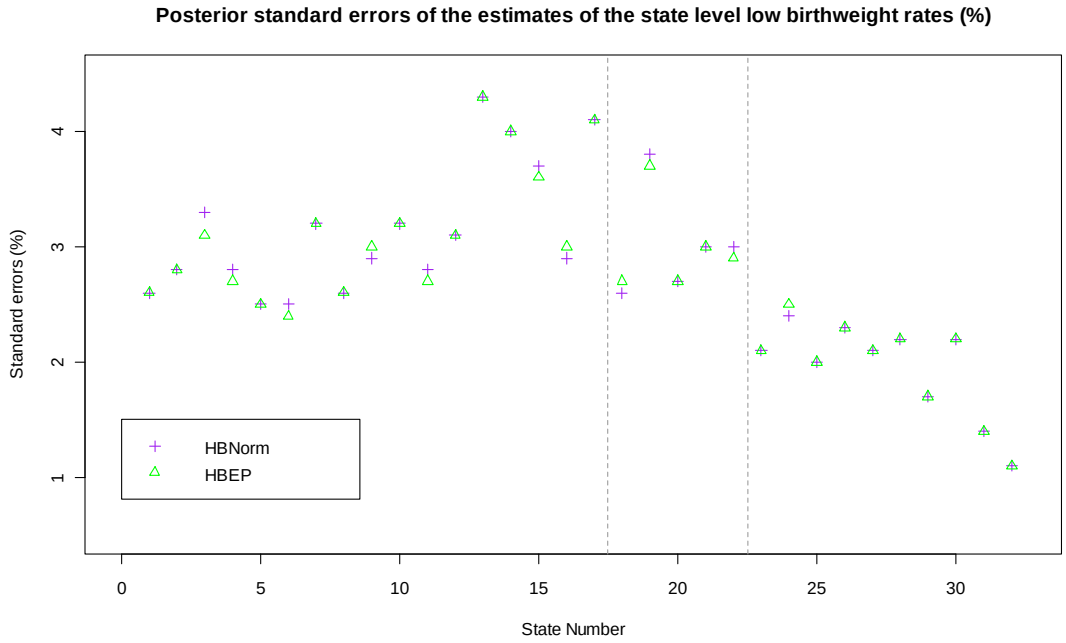

Figure 6-2: Posterior standard errors for the HB estimates of low birthweight rates (in percentage) for the 32 states with samples, where states were sorted by the sample sizes and true proportions P_i.

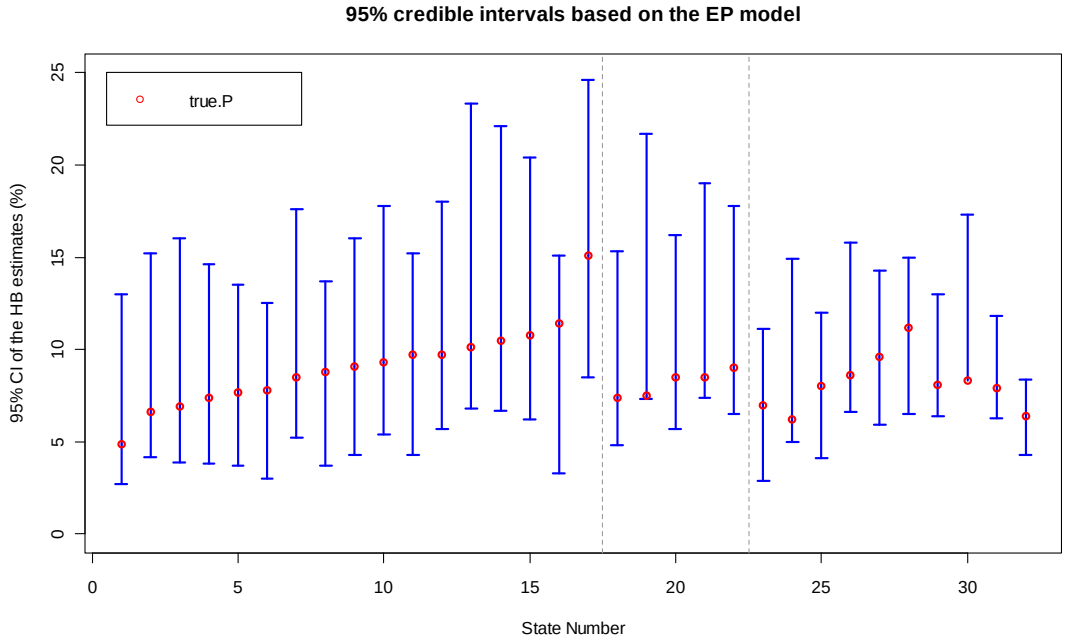

Figure 6-3: 95% credible intervals of the HB estimates of low brithweight rates (in percentages) based on the EP model for the 32 states with samples, where states were sorted by the sample sizes and true proportions P_i (the red dot points are the true P_i).

In order to estimate the P_i for the states without sampled births, we tried the following two approaches:

1) Synthetic approach: using the overall mean to predict for the states without sampled births:

$$\hat{P}_i^{Syn} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{c_i}\sum_{k=1}^{n_{ij}} y_{ijk}}{\sum_{i=1}^{m}\sum_{j=1}^{c_i} n_{ij}},\ i = 1,...,18.$$

The associated standard errors can be estimated using Taylor series linearization method. SUDAAN PROC DESCRIPT was used to obtain both the point estimates and the standard errors. The certainty PSUs were treated as strata instead of PSUs for the variance computation. Note that this synthetic is very simplistic. We consider it here just for comparison purpose. In practice, more complex synthetic approaches are usually preferred. For example, one can group states by geography, urbanicity, etc., and then use group means and/or use some control variables such as race distributions to adjust the overall estimates by those variables.

2) HB approach: using the fully Bayesian approach described in Section 6.3 to predict for states without sampled births.

Figure 6-4a presents the predicted values of the state level low birthweight rates for the 18 states without sampled births based on the two HB models and the synthetic method. For facilitate comparisons, we plot the residuals of the point estimates from the true values in Figure 6-4b. For further analysis, we computed the average of the residuals (AR) of each estimate over the 18 states which was defined

as $AR = \frac{1}{18}\sum_{i=1}^{18}\left[estimate(P_i) - P_i\right]$. The results of AR for HBNorm.P, HBEP.P, and Syn.P are 0.56%, 0.57%, 0.01% respectively, which are all positive, but pretty close to zero.

Figure 6-5 presents the associated standard errors or posterior standard errors of these point estimates. The graphs indicate that both HB models perform better than the synthetic method in predicting P_i for the states with no sampled births. There is again little difference between the estimates produced by the two HB methods.

Figure 6-6 displays the 95% prediction credible intervals based on the *Bernoulli-Logit-EP* model. The performance of the credible intervals looks similar to those in the middle panel of Figure 6-3.

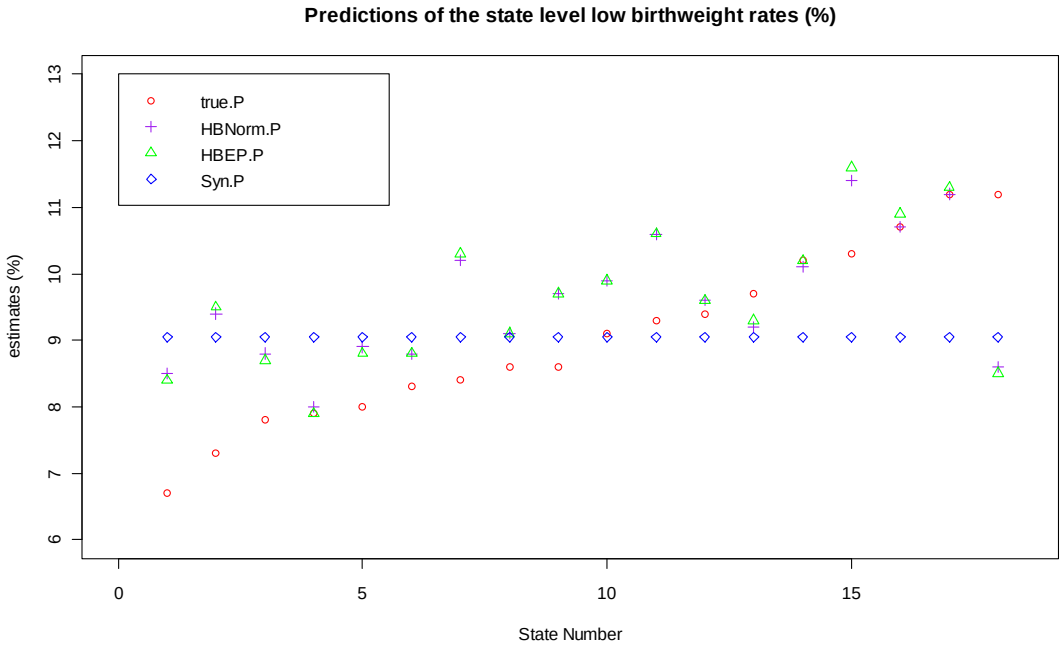

Figure 6-4a: Point estimates for predicting the low birthweight rates (in percentages) for the 18 states with no sampled births, where states were sorted by the true proportions P_i.

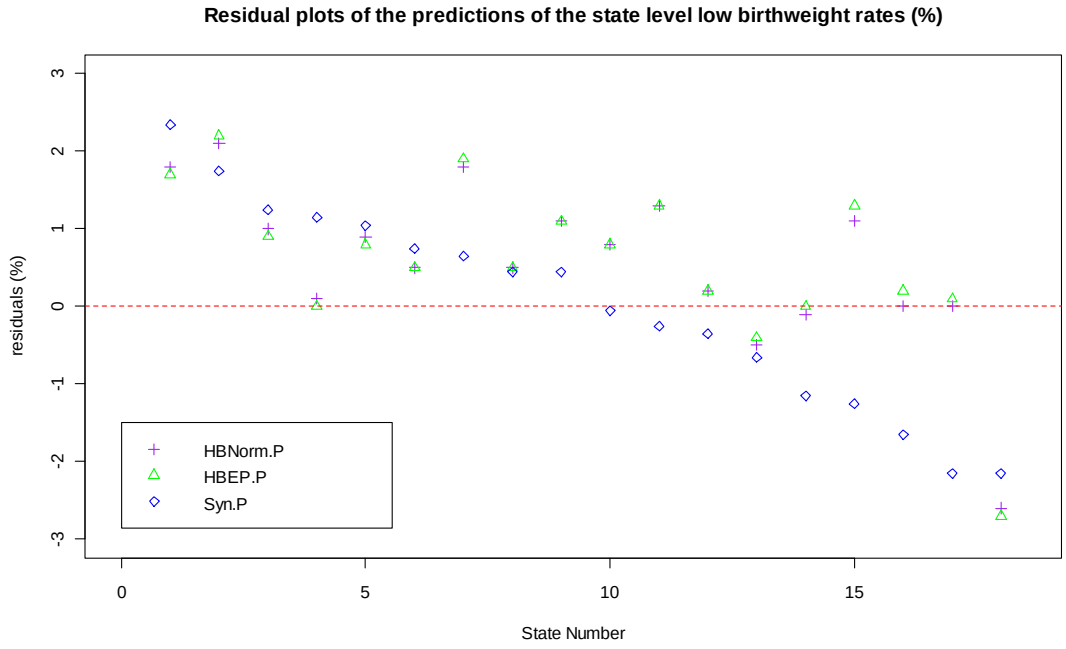

Figure 6-4b: Residual plots for the point estimates for predicting the low birthweight rates (in percentages) for the 18 states with no sampled births, where the residuals were defined by $estimate(P_i) - P_i$, and states were sorted by the true proportions P_i .

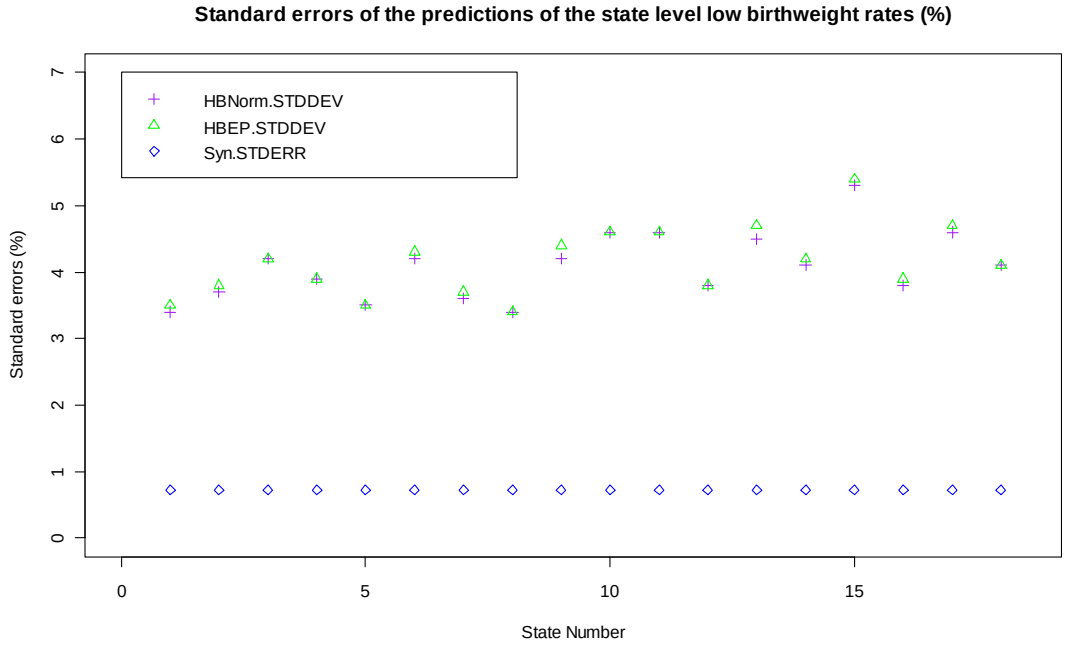

Figure 6-5: Standard errors or posterior standard errors of different point estimates for predicting the low birthweight rates (in percentages) for the 18 states without samples, where states were sorted by the true proportions P_i .

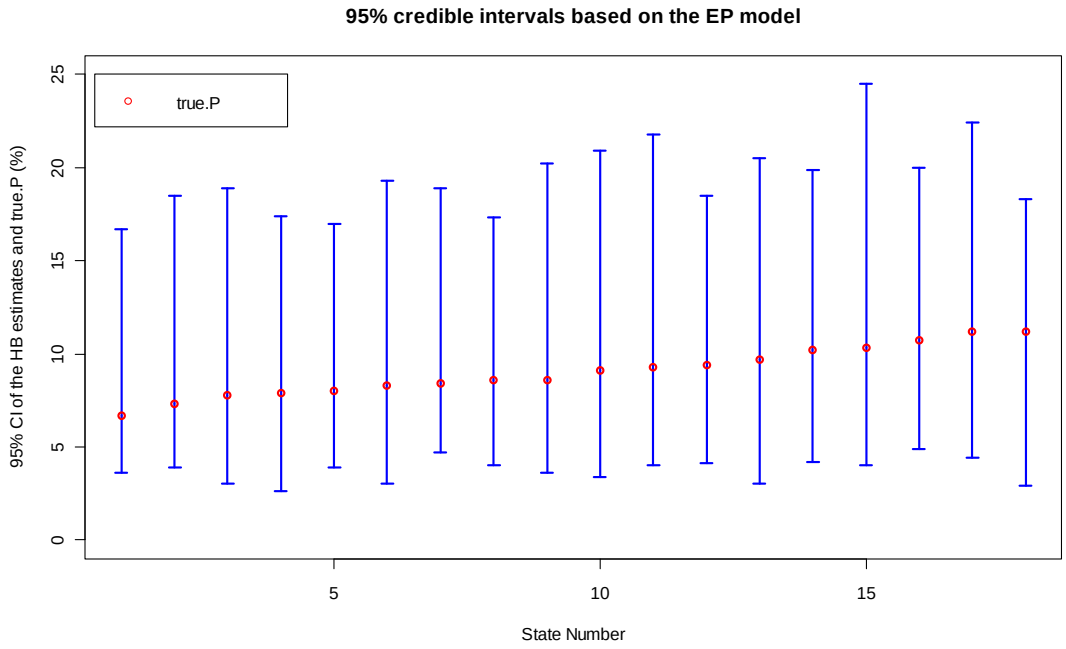

Figure 6-6: 95% credible intervals of the HB estimates based on the EP model for predicting the low birthweight rates (in percentages) for the 18 states with no sampled births, where states were sorted by the true proportions P_i.

It is hard to distinguish the performances between the two HB estimators from the above figures. For further evaluation, we computed the absolute relative deviation $ARD_i = 100\% \times \left| \hat{P}_i - P_i \right| / P_i$ for the point estimates, where $\hat{P}_i$ is a point estimate of P_i, $i = 1,...,50$. We then took average of the ARD_i overall and by three group of states: 1) states with $n_i = 30$; 2) states with $n_i \geq 60$; and 3) states with no sampled births. The average absolute relative deviations are presented in Table 6-1. The first column in the table is for the direct estimates and the second and third columns are for the HB estimates. From the table, we can see that HB estimates perform much better than the direct estimates in terms of ARD. There is no meaningful difference between the two HB models.

Table 6-1: Average absolute relative deviations of the point estimates in estimating P_i (in percentages)

State group	Direct estimate	HB estimate using the Normal Model	HB estimate using the EP Model
Overall	—	15.9	15.9
Small n_i ($n_i = 30$)	57.6	15.6	15.7
Medium to large n_i ($n_i \geq 60$)	41.8	22.6	22.3
Zero n_i ($n_i = 0$)	—	10.7	10.7

We present the posterior means and associated posterior standard errors for the hyperparameters of the two HB models in Table 6-2 below. The estimates for the common hyperparameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$, σ_u and σ_v based on the *Bernoulli-Logit-Normal* model are very close to those based on the *Bernoulli-Logit-EP* model. The posterior means of the two kurtosis parameters (φ_u and φ_v) of *Bernoulli-Logit-EP* model are very close to 0.5, i.e., the normal case. This explains why we can not detect much difference between the two HB models.

Table 6-2: Posterior means and standard deviations of the hyperparameters in the HB models

	HB estimate using the Normal Model		HB estimate using the EP Model	
hyperparameters	Posterior mean	sd	Posterior mean	sd
β_0	-3.76	1.38	-3.97	1.24
β_1	108.30	67.65	112.40	67.10
β_2	3.06	3.36	3.56	3.02
σ_u	0.292	0.150	0.316	0.142
σ_v	0.313	0.159	0.305	0.165
φ_u	—	—	0.466	0.294
φ_v	—	—	0.499	0.289

6.5 Concluding Remarks

In this Chapter, we have extended the *Bernoulli-Logit-EP* model developed in Chapter 3 to a more general model which is suitable for binary data collected through a two-stage sample design. The data analysis showed clear superiority of the hierarchical Bayesian estimates over the direct estimates for estimating small area proportions. However, we did not detect strong evidence showing that the *Bernoulli-Logit-EP* model performs better than the *Bernoulli-Logit-Normal* model for the sample dataset being considered. We need conduct further investigation to find out the reason.

We note that all the credible intervals of the HB estimates cover the true values, which is too conservative comparing with the nominal 5 percent noncoverage. However, we cannot draw a firm conclusion about the coverage because the data analysis was based only on one sample. We might expect the credible intervals of 2 states out the 50 states to fail to cover the true P_i.

For simplicity, the data was assumed free of nonsampling errors. However, this assumption can be easily violated for real surveys due to nonresponse. Further, all the sampled births were drawn independent within PSU based on our design. In practice, over-sampling for minority groups, clustering by minority and socioeconomic status within county, etc. are often used in a complex multistage survey. A more complex model and further research are needed to account for such situations.

Appendix for Chapter 6

Appendix A: Full conditional distributions for the two HB models

A.1 *Bernoulli-Logit-Normal* model

Assume the following prior assumptions for the hyperparameters:

$$f(\boldsymbol{\beta}) \propto 1\ ; \quad \sigma_u \sim Uniform(0,\ L)\ ; \quad \sigma_v \sim Uniform(0,\ L) \tag{A.1}$$

The full conditional distributions for all the model parameters of the HB version of the *Bernoulli-Logit-Normal* model are as follows:

$$\theta_{ij} \mid \mathbf{y}_s, \mathbf{v}, \boldsymbol{\beta}, \sigma_v^2, \sigma_u^2 \propto \begin{cases} \theta_{ij}^{y_{ij}-1}(1-\theta_{ij})^{n_{ij}-y_{ij}-1} \exp\left\{-\dfrac{\left[\text{logit}(\theta_{ij}) - \mathbf{x}_{ij}'\boldsymbol{\beta} - v_i\right]^2}{2\sigma_u^2}\right\}, & \text{for } j \in s_i,\ j = 1,...,c_i \\ \dfrac{1}{\theta_{ij}(1-\theta_{ij})} \exp\left\{-\dfrac{\left[\text{logit}(\theta_{ij}) - \mathbf{x}_{ij}'\boldsymbol{\beta} - v_i\right]^2}{2\sigma_u^2}\right\}, & \text{for } j \notin s_i \end{cases}$$

where $0 < \theta_{ij} < 1,\ \ i = 1,...,m$;

$$v_i \mid \mathbf{y}_s, \boldsymbol{\theta}, \boldsymbol{\beta}, \sigma_v^2, \sigma_u^2 \sim N\left[\frac{\sigma_v^2 \sum_{j=1}^{c_i}\left(\text{logit}(\theta_{ij}) - \mathbf{x}_{ij}'\boldsymbol{\beta}\right)}{c_i\sigma_v^2 + \sigma_u^2}, \frac{\sigma_v^2\sigma_u^2}{c_i\sigma_v^2 + \sigma_u^2}\right], \text{ for } v_i \in R,$$

$i = 1,...,m$;

$$\boldsymbol{\beta} \mid y_s, \boldsymbol{\theta}, \mathbf{v}, \sigma_v^2, \sigma_u^2 \sim N\left[\left(\sum_{i=1}^{m}\sum_{j=1}^{c_i}\mathbf{x}_{ij}\mathbf{x}_{ij}'\right)^{-1}\left[\sum_{i=1}^{m}\sum_{j=1}^{c_i}\mathbf{x}_{ij}'\left(\text{logit}(\theta_{ij}) - v_i\right)\right],\ \sigma_u^2\left(\sum_{i=1}^{m}\sum_{j=1}^{c_i}\mathbf{x}_{ij}\mathbf{x}_{ij}'\right)^{-1}\right],$$

for $\boldsymbol{\beta} \in R$;

$$\sigma_u^2 \mid \mathbf{y}_s, \boldsymbol{\theta}, \mathbf{v}, \boldsymbol{\beta}, \sigma_v^2 \sim ING\left[\frac{1}{2}\sum_{i=1}^{m} c_i - 1,\ \ \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{c_i}\left(\text{logit}(\theta_{ij}) - \mathbf{x}_{ij}'\boldsymbol{\beta} - v_i\right)^2\right],$$

for $\sigma_u \in (0, L)$;

$$\sigma_v^2 \mid \mathbf{y}_s, \boldsymbol{\theta}, \mathbf{v}, \boldsymbol{\beta}, \sigma_u^2 \sim ING\left[\frac{1}{2}m - 1, \quad \frac{1}{2}\sum_{i=1}^{m} v_i^2\right], \text{ for } \sigma_v \in (0, L).$$

A.2 *Bernoulli-Logit-EP* model

In addition to the prior assumptions (A.1) used for the *Bernoulli-Logit-Normal* model, we assume the following prior assumptions for the hyperparameters φ_u and φ_v:

$$\varphi_u \sim Uniform(0,\ 1); \qquad \varphi_v \sim Uniform(0,\ 1) \tag{A.2}$$

The full conditional distributions for all the model parameters of the HB version of the *Bernoulli-Logit-EP* model are as follows:

$$\theta_{ij} \mid \mathbf{y}_s, \mathbf{v}, \boldsymbol{\beta}, \sigma_v^2, \sigma_u^2, \varphi_u, \varphi_v$$

$$\propto \begin{cases} \theta_{ij}^{y_{ij}-1}(1-\theta_{ij})^{n_{ij}-y_{ij}-1} \exp\left[-\dfrac{(c_{0u})^{1/(2\varphi_u)}\left|\text{logit}(\theta_{ij}) - \mathbf{x}_{ij}'\boldsymbol{\beta} - v_i\right|^{1/\varphi_u}}{\sigma_u^{1/\varphi_u}}\right], & \text{for } j \in s_i,\ j = 1, \ldots, c_i \\ \dfrac{1}{\theta_{ij}(1-\theta_{ij})} \exp\left[-\dfrac{(c_{0u})^{1/(2\varphi_u)}\left|\text{logit}(\theta_{ij}) - \mathbf{x}_{ij}'\boldsymbol{\beta} - v_i\right|^{1/\varphi_u}}{\sigma_u^{1/\varphi_u}}\right], & \text{for } j \in s_i^c \end{cases}$$

where $0 < \theta_{ij} < 1, \quad i = 1, \ldots, m$;

$$v_i \mid \mathbf{y}_s, \boldsymbol{\theta}, \boldsymbol{\beta}, \sigma_v^2, \sigma_u^2, \varphi_u, \varphi_v$$

$$\propto \exp\left[-\frac{(c_{0u})^{1/(2\varphi_u)}\sum_{j=1}^{c_i}\left|\text{logit}(\theta_{ij}) - \mathbf{x}_{ij}'\boldsymbol{\beta} - v_i\right|^{1/\varphi_u}}{\sigma_u^{1/\varphi_u}} - \frac{(c_{0v})^{1/(2\varphi_v)}\left|v_i\right|^{1/\varphi_v}}{\sigma_v^{1/\varphi_v}}\right],$$

for $v_i \in R$, $i = 1,...,m$;

$$\boldsymbol{\beta} \mid \mathbf{y}_s, \boldsymbol{\theta}, \mathbf{v}, \sigma_v^2, \sigma_u^2, \varphi_u, \varphi_v \propto \exp\left[-\frac{(c_{0u})^{1/(2\varphi_u)} \sum_{i=1}^{m} \sum_{j=1}^{c_i} \left| \text{logit}(\theta_{ij}) - \mathbf{x}_{ij}'\boldsymbol{\beta} - v_i \right|^{1/\varphi_u}}{\sigma_u^{1/\varphi_u}} \right],$$

for $\beta \in R$;

$$\sigma_u \mid \mathbf{y}_s, \boldsymbol{\theta}, \mathbf{v}, \boldsymbol{\beta}, \sigma_v^2, \varphi_u, \varphi_v$$

$$\propto \frac{1}{(\sigma_u)^{\sum_{i=1}^{m} c_i}} \exp\left\{ -\frac{(c_{0u})^{1/(2\varphi_u)} \sum_{i=1}^{m} \sum_{j=1}^{c_i} \left| \text{logit}(\theta_{ij}) - \mathbf{x}_{ij}'\boldsymbol{\beta} - v_i \right|^{1/\varphi_u}}{\sigma_u^{1/\varphi_u}} \right\}, \text{ for } \sigma_u \in (0, L);$$

$$\sigma_v \mid \mathbf{y}_s, \boldsymbol{\theta}, \mathbf{v}, \boldsymbol{\beta}, \sigma_u, \varphi_u, \varphi_v \propto \frac{1}{(\sigma_v)^m} \exp\left[-\frac{(c_{0v})^{1/(2\varphi_v)} \sum_{i=1}^{m} |v_i|^{1/\varphi_v}}{\sigma_v^{1/\varphi_v}} \right], \text{ for } \sigma_v \in (0, L);$$

$$\varphi_u \mid \mathbf{y}_s, \boldsymbol{\theta}, \mathbf{v}, \boldsymbol{\beta}, \sigma_u, \sigma_v, \varphi_v$$

$$\propto (c_{1u})^{\sum_{i=1}^{m} c_i} \exp\left[-\frac{(c_{0u})^{1/(2\varphi_u)} \sum_{i=1}^{m} \sum_{j=1}^{c_i} \left| \text{logit}(\theta_{ij}) - \mathbf{x}_{ij}'\boldsymbol{\beta} - v_i \right|^{1/\varphi_u}}{\sigma_u^{1/\varphi_u}} \right], \text{ for } 0 < \varphi_u < 1;$$

where $c_{0u} = \Gamma(3\varphi_u)/\Gamma(\varphi_u)$, $c_{1u} = \sqrt{c_{0u}} / \left(2\varphi_u \Gamma(\varphi_u)\right)$;

$$\varphi_v \mid \mathbf{y}_s, \boldsymbol{\theta}, \mathbf{v}, \boldsymbol{\beta}, \sigma_u, \sigma_v, \varphi_u \propto (c_{1v})^m \exp\left[-\frac{(c_{0v})^{1/(2\varphi_v)} \sum_{i=1}^{m} |v_i|^{1/\varphi_v}}{\sigma_v^{1/\varphi_v}} \right], \text{ for } 0 < \varphi_u < 1;$$

where $c_{0v} = \Gamma(3\varphi_v)/\Gamma(\varphi_v)$, $c_{1v} = \sqrt{c_{0v}} / \left(2\varphi_v \Gamma(\varphi_v)\right)$.

Appendix B: WinBUGS Code for the two HB models

B.1 For the *Bernoulli-Logit-Normal* model

```
model
{
#L1 number of sampled PSUs
for (i in 1:L1)
{
yobs[i] ~ dbin(theta[i], n[i])
logit(theta[i])<- inprod(beta[], Xs[i,])+v[stateID[i]]+u[i]
y2[i]~ dbin(theta[i], N2[i])
}
#L3 number of nonsampled PSUs
for (i in 1:L3) {
logit(theta3[i])<-inprod(beta[], Xns[i,])+v[stateID3[i]]+u3[i]
y3[i] ~ dbin(theta3[i], N3[i])
}
#m states
for (j in 1:m)
{
v[j]~dnorm(0, precisonv)
}
for (i in 1:L1)
{
u[i]~dnorm(0, precisonu)
}
for (i in 1:L3)
{
u3[i]~dnorm(0, precisonu)
}
for ( i in 1:p)
{
beta[i]~dflat()
}
precisonu<-1/sig2u
precisonv<-1/sig2v
sig2v<-pow(sigmav, 2)
sig2u<-pow(sigmau, 2)
sigmav~dunif(0, 100)
sigmau~dunif(0, 100)
}
```

B.2 For the *Bernoulli-Logit-EP* model

```
model
{
#L1 number of sampled PSUs
```

```
for (i in 1:L1)
{
yobs[i] ~ dbin(theta[i], n[i])
logit(theta[i])<- inprod(beta[], Xs[i,])+v[stateID[i]]+u[i]
y2[i]~ dbin(theta[i], N2[i])
}
#L3 number of nonsampled PSUs
for (i in 1:L3)
{
logit(theta3[i])<-inprod(beta[], Xns[i,])+v[stateID3[i]]+u3[i]
y3[i] ~ dbin(theta3[i], N3[i])
}
#m states
#trick for specifying EP priors for v[j]
C<-10000
for (j in 1:m)
{
zerov[j]<-0
v[j]~dunif(-10000,10000)
phiv[j]<- -(log(c1v)-log(sigmav)-
pow(abs(sqrt(c0v)*v[j]/sigmav),1/psiv))+C
zerov[j]~dpois(phiv[j])
}
c0v<-exp(loggam(3*psiv))/exp(loggam(psiv))
c1v<-sqrt(c0v)/(2*psiv*exp(loggam(psiv)))
#trick for specifying EP priors for u[i]
for (i in 1:L1)
{
zerou[i]<-0
u[i]~dunif(-10000,10000)
phiu[i]<- -(log(c1u)-log(sigmau)-
pow(abs(sqrt(c0u)*u[i]/sigmau),1/psiu))+C
zerou[i]~dpois(phiu[i])
}
c0u<-exp(loggam(3*psiu))/exp(loggam(psiu))
c1u<-sqrt(c0u)/(2*psiu*exp(loggam(psiu)))
#trick for specifying EP priors for u[i]
for (i in 1:L3)
{
zerou3[i]<-0
u3[i]~dunif(-10000,10000)
phiu3[i]<- -(log(c1u)-log(sigmau)-
pow(abs(sqrt(c0u)*u3[i]/sigmau),1/psiu))+C
zerou3[i]~dpois(phiu3[i])
}
# end of trick
for ( i in 1:p)
{
beta[i]~dflat()
}
psiv~dunif(0,1)
psiu~dunif(0,1)
```

```
sigmav~dunif(0, 100)
sigmau~dunif(0, 100)
sig2v<-pow(sigmav, 2)
sig2u<-pow(sigmau, 2)
}
```

Chapter 7: Summary and Future Research

In this dissertation, we have developed new statistical methods that are useful for estimating small area proportions using survey data. Throughout the dissertation, our main goal has been to develop new small area models that incorporate non-normality and non-linearity for dichotomous variables under different complex sampling designs, and to demonstrate how to make inferences using the new models. We have considered both area level and unit level models for an unclustered population, and also considered unit level models for a clustered population under a two-stage sample design.

We first explored alternatives to the well-known Fay-Herriot model. The proposed beta-logistic model has three advantages over the Fay-Herriot model to deal with survey-weighted proportions: 1) it assumes a beta distribution instead of a normal distribution for the sampling model to deal with potential asymmetry or skewness of the sampling distribution; 2) it utilizes a *logit* link at the linking model to take care of nonlinear cases and to guarantee the estimates fall in the (0, 1) range; and 3) it assumes the sampling variances are unknown and can be estimated simultaneously through the HB estimation process. The simulation results indicated that the beta-logistic model has fair coverage properties though with large simulation error.

However, zero survey-weighted proportions frequently occurred in the repeatedly simulated sample data, especially in areas with small sample sizes. We discovered that the weakness of the proposed beta-logistic model was its inherent incapability to deal with zero direct estimates. The zero direct estimates were

converted to very small positive numbers in that study in order to obtain results. Nevertheless, the conversion may introduce addition bias and lead to invalid inferences. Consequently, we may fix the zero problems by improving the original model. One approach is to develop a two-part model, which assumes two modeling stages for the direct survey estimates: one determining whether the direct estimate is zero and the other determining the actual measure of parameters of interest if it is non-zero.

In the second part of the dissertation research, we proposed robust unit level mixed models by assuming a class of distributions which includes normality for the random effects as a special case. These models were developed under a single stage sample design and a two-stage sample design. We explored hierarchical Bayesian inferences using different approximate methods including MCMC, first- and second-order Laplace approximation, Gauss-Hermite Quadrature integration, and Monte Carlo integration methods, and demonstrated the advantages of MCMC especially under non-normal cases. We have also studied the empirical best prediction approach for the model developed under a single stage sample design. We prefer the hierarchical Bayesian approach since it is more flexible and performs well compared to EBP.

The class of distributions considered can capture kurtosis in the distribution of the random effects. However, other nonnormality phenomena such as skewness have not been studied. We will study a more general class of distributions which can capture both skewness and kurtosis and incorporate it in our robust unit level mixed models. In addition, in practice, large-scale national surveys often employ complex

multi-stage survey designs involving several layers of stratification and clustering. We plan to generalize our proposed unit level models to incorporate complex multi-stage designs.

Bibliography

Australian Bureau of Statistics (2005), "A guide to small area estimation," available at *http://www.nss.gov.au/nss/home.NSF/pages/Small+Areas+Estimates?OpenDocument*.

Arora, V., and Lahiri, P. (1997), "On the superiority of the Bayesian method over the BLUP in small area estimation problems," *Statistical Sinica*, 7, 1053-1063.

Arora, V., Lahiri, P., and Mukherjee, K. (1997), "Empirical Bayes estimation of finite population means from complex surveys," *Journal of the American Statistical Association*, 92, 1555-1562.

Asher, J., and Fisher, R. (2000), "Alternative scaling parameter functions in a hierarchical Bayes model of U.S. county poverty rates," *Proceedings of the American Statistical Association*, 283-288.

Azzalini, A. (1985), "A class of distributions which includes the normal ones," *Scandinavian Journal of Statistica*, 12, 171-178.

Azzalini, A. (1986), "Further results on a class of distributions which include the normal ones," *Statistica*, 46, 199-208.

Battese, G.E., Harter, R.M., and Fuller. W.A. (1988), "An error-components model for prediction of county crop areas using survey and satellite data," *Journal of the American Statistical Association*, 80, 28-36.

Bolfarine, H., and Zacks, S. (1992), *Prediction Theory for Finite Populations*, New York: Springer-Verlag.

Box, G.E.P., and Tiao, G.C. (1973), *Bayesian Inference in Statistical Analysis*, Reading, MA: Addison-Wesley.

Bradley, R.A., and Gart, J.J. (1962), "The asymptotic properties of ML estimators when sampling from associated populations," *Biometrika*, 49, 205–214.

Brewer, K.R.W. (1963), "Ratio Estimation and Finite Populations: Some results deducible from the assumption of an underlying stochastic process," *Australian Journal of Statistics*, 5, 93-105.

Butar, F.B., and Lahiri, P. (2002), "Empirical Bayes estimation of several population means and variances under random sampling variances model," *Journal of Statistical Planning and Inference*, 102, 59-69.

Butar, F.B., and Lahiri, P. (2003), "On measure of uncertainty of empirical Bayes small area estimators," *Journal of Statistical Planning and Inference*, 112, 63-76.

Carter, G.M., and Rolph, J.E. (1974), "Empirical Bayes methods applied to estimating fire alarm probabilities," *Journal of the American Statistical Association*, 69, 880-885.

Chatterjee, S., and Lahiri, P. (2008), "A simple computation method for estimating mean squared prediction error in general small-area model," *Proceedings of the American Statistical Association*.

Chatterjee, S., Lahiri, P., and Li, H. (2008), "Parametric bootstrap approximation to the distribution of EBLUP, and related prediction intervals in linear mixed models," *Annals of Statistics*, 36, 1221-1245.

Chen, M., Shao, Q., and Ibraham, J.G. (2000), *Monte Carlo Methods in Bayesian Computation*, New York: Springer-Verlag.

Chen, S.J. (2001), "Empirical best prediction and hierarchical Bayes methods in small area estimation", Ph.D. Dissertation, Dept. of Mathematics and Statistics, University of Nebraska-Lincoln.

Chib, S., and Greenberg, E. (1995), "Understanding the Metropolis-Hastings algorithm," *The American Statistician*, 49, 327-335.

Choy, S.T.B., and Smith, A.F.M. (1997a), "On robust analysis of a normal location parameter," *Journal of Royal Statistical Society*, *B*, 59, 463-474.

Choy, S.T.B., and Smith, A.F.M. (1997b), "Hierarchical models with scale mixtures of normal distributions," *TEST*, 6, 202-221.

Christiansen, C.L., and Morris C.N. (1997), "Hierarchical Poisson regression modeling," *Journal of the American Statistical Association*, 92, 618-32.

Citro, C., and Kalton, G. (Eds.) (2000), *Small-Area Income and Poverty Estimates: Priorities for 2000 and Beyond*, Washington, DC: National Academy Press.

Cochran, W. G. (1939), "The use of analysis of variance in enumeration by sampling," *Journal of the American Statistical Association*, 34, 492-510.

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

Datta, G.S., Fay, R.E., and Ghosh, M. (1991), "Hierarchical and empirical Bayes multivariate analysis in small area estimation," *Proceedings of Bureau of the Census Annual Research Conference*, U.S. Bureau of the Census, Washington, DC, 63-79.

Datta, G.S., and Lahiri, P. (1995), "Robust hierarchical Bayes estimation of small area characteristics in the presence of covariates and outlier," *Journal of Multivariate Analysis*, 54, 310-328.

Datta, G.S., and Lahiri, P. (2000), "A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems," *Statistica Sinica*, 10, 613-627.

Datta, G.S., Lahiri, P., and Maiti, T. (2002), "Empirical Bayes estimation of median income of four-person families by state using time series and cross-sectional data," *Journal of Statistical Planning and Inference*, 102, 83-97.

Datta, G.S., Lahiri, P., Maiti, T., and Lu, K.L. (1999), "Hierachical Bayes estimation of the unemployment rates for the stats of the U.S.," *Journal of the American Statistical Association*, 94, 1074-1082.

Datta, G.S., and Ghosh, M (1991), "Bayesian prediction in linear models: Application to small area estimation," *Annals of Statistics*, 19, 1748-1770.

Davis, P.J., and Rabinowitz, P. (1975), *Methods of Numerical Integration*, New York: Academic Press.

Dempster, A. P., and Tomberlin, T. J. (1980), "The analysis of census undercount from a postenumeration survey," *Proceedings of the Conference on Census Undercount*, Arlington, VA, 88-94.

Efron, B., and Morris, C.N. (1975), "Data analysis using Stein's estimator and its generalizations," *Journal of the American Statistical Association*, 70, 311-319.

Erdelyi, A. (1956), *Asymptotic Expansions*, New York: Dover.

Ericson, W.A. (1969), "Subjective Bayesian models in sampling finite population (with discussions)," *Journal of the Royal Statistical Society*, 31, 195-233.

Fabrizi, E., and Trivisano, C. (2007), "Robust models for mixed effects in linear mixed models applied to small area estimation," Submitted to *Journal of Statistical Planning and Inference*.

Farrell, P.J., MacGibbon, B., and Tomberlin, T.J. (1994), "Protection against outliers in empirical Bayes estimation," *Canadian Journal of Statistics*, 22, 365-376.

Farrell, P.J., MacGibbon, B., and Tomberlin, T.J. (1997a), "Empirical Bayes estimators of small area proportions in multistage designs," *Statistical Sinica*, 7, 1065-1083.

Farrell, P.J., MacGibbon, B., and Tomberlin, T.J. (1997b), "Empirical Bayes small-area estimation using logistic regression models and summary statistics," *Journal of Business & Economic Statistics*, 15, 101-108.

Fay, R.E. (1987), "Application of Multivariate Regression to Small Domain Estimation," in Platek, R., Rao, J.N.K., Sarndal, C.E., and Sing, M.P. (eds.), *Small Area Statistics*, New York: Wiley, 91-102.

Fay, R.E., and Herriot, R.A. (1979), "Estimates of income for small places: An application of James-Stein procedure to census data," *Journal of the American Statistical Association*, 74, 269-277.

Fisher, R., and Asher, J. (1999), "Bayesian hierarchical modeling of U.S. county poverty rates," *Proceedings of the American Statistical Association*.

Gelman, A. (2006), "Prior distributions for variance parameters in hierarchical models," *Bayesian Analysis*, 1, 515-533.

Gelman, A., Carlin, J.B., Carlin, Stern, H.S., and Rubin, D.B. (2004), *Bayesian Data Analysis*, 2nd edition, Chapman & Hall/CRC.

Gelman, A., and Rubin, D.B. (1992), "Inference from iterative simulation using multiple sequence," *Statistical Science*, 7, 457-472.

Gelfand, A. E., and Smith, A.F.M. (1990), "Sampling-based approaches to calculating marginal densities," *Journal of the American Statistical Association*, 85, 269-277.

Gershunskaya, J., Jiang, J., and Lahiri, P. (2008). "Resampling methods in surveys", in Rao, C.R., and Pfeffermann, D. (eds.), *Sample Surveys: Theory Methods and Inference*, in press.

Ghosh, M. and Lahiri, P. (1987a), "Robust empirical Bayes estimation of means from stratified samples," *Journal of the American Statistical Association*, 82, 1153-1162.

Ghosh, M., and Lahiri, P. (1987b), "Robust empirical Bayes estimation of variances from stratified samples," *Sankhya*, ***B*****,** 49, 78-89.

Ghosh, M., and Lahiri, P. (1988), "Bayes and empirical Bayes analysis in multistage sampling," *Statistical Decision Theory and Related Topics IV*, 1, 195-212.

Ghosh, M., Lahiri, P., and Tiwari, R.C. (1989), "Nonparametric Bayes and empirical Bayes estimation of the distribution function and the mean," *Communications in Statistics: Theory and Methods*, 18, 121-46.

Ghosh, M., and Meeden, G. (1986), "Empirical Bayes estimation in finite population sampling," *Journal of the American Statistical Association*, 81, 1058-1062.

Ghosh, M., and Meeden, G. (1997), *Bayesian Methods for Finite Population Sampling*, New York: Chapman and Hall.

Gilks, W.R., Richardson, S., and Spiegelhalter, D.J. (1996), *Markov Chain Monte Carlo in Practice: Interdisciplinary Statistics*, London: Chapman & Hall.

Gnanadesikan, R. (1977), *Methods for Statistical Analysis of Multivariate Observations*, Wiley-Interscience.

Godfrey, K.M. and Barker, D.J. (2000), "Fetal nutrition and adult disease," *American Journal of Clinical Nutrition*, 71, 1344S-1352s.

Gonzales, M.E. (1973), "Use and evaluation of synthetic estimation," *Proceedings of the American Statistical Association*, Social Statistics Section, 33-36.

Gonzales, M.E., and Hoza, C. (1978), "Small-area estimation with application to unemployment and housing estimations," *Journal of the American Statistical Association*, 73, 7-15.

Graubard, B., and Korn, E. (2002), "Inference for superpopulation parameters using sample surveys," *Statistical Science*, 17, 73-96.

Hall, P. and Maiti, T. (2006), "On parametric bootstrap methods for small area estimation," *Journal of Royal Statistical Society*, *B*, 68, 221-238.

Hajek, J. (1971), "Comment," in Godambe, V.P., and D.A. Sprott, D.A. (eds.), *Foundations of Statistical Inference*, Toronto: Holt, Rinehart and Winston, p.236.

Hampel, F. R., Ronchetti, E.M., Rousseeuw, P.J., and Stahel, W. A. (1986), *Robust Statistics: The Approach Based on Influence Functions*, New York: John Wiley and Sons.

Hanurav, T. V. (1967), "Optimal utilization of auxiliary information in π ps sampling of two units from a stratum," *Journal of the Royal Statistical Society*, *B*, 29, 374-391.

He, Z., and Sun, D. (1998), "Hiearchical Bayes estimation of hunting success rates," *Environmental and Ecological Statistics*, 5, 223-236.

He, Z., and Sun, D. (2000), "Hiearchical Bayes estimation of hunting success rates with spatial correlations," *Biometrics*, 56, 360-367.

Hinrichs, P.E. (2003), "Consumer expenditure estimation incorporating generalized variance functions in hierarchical Bayes models", Ph.D. Dissertation, Dept. of Mathematics and Statistics, University of Nebraska-Lincoln.

Hogg, R.V. (1974), "Adaptive robust procedures: A partial review and some suggestions for future applications and theory," *Journal of the American Statistical Association*, 69, 909-923.

Huang, E. T., and Bell, W. R (2006), "Using the t-distribution in small area estimation: An application to SAIPE State poverty model," *Proceedings of the American Statistical Association*, 3142-3149.

Humphrey, C., and Elford, J. (1988), "Social class differences in infant mortality: the problem of competing hypothesis," *Journal of Biosocial Science*, 20, 497-504.

Jiang, J., and Lahiri, P. (2001), "Empirical best prediction for small area inference with binary data," *Annals of Institute of Statistical Mathematics*, 53, 217-243.

Jiang, J., Lahiri, P., and Wan, S. M. (2002), "A unified jackknife theory for empirical best prediction with M-estimation," *Annals of Statistics*, 30, 1782-1810.

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

Jiang, J., and Lahiri, P. (2006b), "Estimation of finite population domain means: A model-assisted empirical best prediction approach," *Journal of the American Statistical Association*, 101, 301-311.

Jiang, J., Rao, J.S., Gu, Z., and Nguyen, T. (2008), "Fence methods for mixed model selection," *Annals of Statistics*, 36, 1669-1692.

Kass, R.E., Tierney, L., and Kadane, J.B. (1988), "Asymptotics in Bayesian computation," in Bernardo, J.M., DeGroot, M.H., D.V. Lindley, D.V., and Smith, A.F.M. (eds.), *Bayesian Statistics*, Oxford University Press. 3, 261-278.

Kass, R.E., and Steffey, D. (1989), "Approximate Bayesian inference in conditionally independent hierarchical models (Parametric empirical Bayes models)," *Journal of the American Statistical Association*, 84, 717-726.

Kish, L. (1965), *Survey sampling*, New York: John Wiley and Sons.

Kleffe, J., and Rao, J.N.K. (1992), "Estimation of mean square error of empirical best linear unbiased predictors under a random error variance linear model," *Journal of Multivariate Analysis*, 43, 1-15.

Kott, P. (1989), "Robust small domain estimation using random effects modeling," *Survey Methodology*, 15, 3-12.

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

Lahiri, P., and Tiwari, R.C. (1991), "Nonparametric Bayes and empirical Bayes estimation of variances from stratified samples," *Sankhya*, *B*, 52, 105-118.

Lahiri, P., and Rao, J.N.K. (1995), "Robust estimation of mean squared error of small area estimators," *Journal of the American Statistical Association*, 82, 758-766.

Laplace, P.S. De (1847), *Oeuvres 7*, Paris: Imprimerie Royale.

Lepage, G.P. (1978), "A new algorithm for adaptive multidimensional integration," *Journal of Computational Physics*, 27, 192-203.

Li, H. (2007), "Small area estimation: an empirical best linear unbiased prediction approach," Ph.D. Dissertation, Dept. of Mathematics and Statistics, University of Maryland.

Li, Y, and Lahiri, P. (2007), "Robust model-based and model-assisted predictors of the finite population total," *Journal of the American Statistical Association*, 102, 664-673.

Little, R. (1983), "Estimating a finite population mean from unequal probability samples," *Journal of the American Statistical Association*, 78, 596-604.

Little, R. (1993), "Post-stratification: A modeler's perspective," *Journal of the American Statistical Association*, 88, 1001-1012.

Little, R. (2004), "To model or not to model? Competing modes of inference for finite population sampling," *Journal of the American Statistical Association*, 99, 546-556.

Liu, Q., and Pierce, D.A. (1994), "A note on Gauss-Hermite Quadrature," *Biometrika*, 81, 624-629.

Lohr, S. (1999), *Sampling, Design and Analysis*, Pacific Grove; CA: Duxbury.

Lohr, S., and Rao, J.N.K. (2007), "Jackknife estimation of mean square error of small area predictors in nonlinear mixed models", Submitted for publication.

MacGibbon, B., and Tomberlin, T.J. (1989), "Small area estimates of proportions via empirical Bayes techniques," *Survey Methodology*, 15, 237-252.

Malec, D., Davis, W., and Cao, X. (1999), "Small area estimates of overweight prevalence using sample selection adjustment," *Statistics in Medicine*, 18, 3189-3200.

Malec, D., and Sedransk, J. (1985), "Bayesian inference for finite population in multistage cluster samping," *Journal of the American Statistical Association*, 80, 897-902.

Malec, D., Sedransk, J., Moriarity, C.L., and LeClere, F.B. (1997), "Small area inference for binary variables in National Health Interview Survey," *Journal of the American Statistical Association*, 92, 815-826.

Malec, D., Sedransk, J. and Tompkins, L. (1993), "Bayesian predictive inference for small areas for binary variables in the national health interview survey," in Gatsonis, C., Hodges, J.S., Kass, R.E., and Singpurwalla, N.D. (eds.), *Case Studies in Bayesian Statistic*, New York: Springer Verlag, 377-389.

Maples, J., and Bell, W.R. (2005). "Evaluation of school district poverty estimates: Predictive models using IRS income tax data," *Proceedings of the American Statistical Association*.

McCulloch, C.E. (2003), *Generalized Linear Mixed Models*, NSF-CBMS Regional Conference Series in Probability and Statistics, 7. Beachwood, OH: Institute of Mathematical Statistics.

Meeden, G. (1999), "A noninformative Bayesian approach for two-stage cluster sampling," *Sankhya*, *B*, 61, 133-144.

Mohadjer, L., Rao, J.N.K., Liu, B., Krenzke, T., and Van de Kerckhove, W. (2007), "Hierarchical Bayes small area estimates of adult literacy using unmatched sampling and linking models," *Proceedings of the American Statistical Association*, Section on Survey Research Methods.

Mohadjer, L., Kalton, G., Krenzke, T., Liu, B., Van de Kerckhove, W., Li, L., Sherman, D., Dillman, J., Rao, J., and White, S. (2008), *National Assessment of Adult Literacy: Indirect County and State Estimates of the Percentage of Adults at the Lowest Level of Literacy for 1992 and 2003* (*NCES 2008-467*), National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education, Washington, D.C.

Morris, C.N. (1983), "Parametric empirical Bayes inference: theory and applications," *Journal of the American Statistical Association*, 78, 47-59.

Morris, C.N. (1988), "Approximating posterior distributions and posterior moments," in Bernardo, J.M., Degroot, M.H., Lindley, D.V., and Smith, A.F.M. (eds.), *Bayesian Statistics*, Oxford University Press, 3, 327-344.

Morris, C.N. (2006), "Discussion of mixed model prediction and small area estimation," *Test*, 15, 1-96.

Morris, C.N., and Christiansen, C.L. (1994), "Hierarchical models for ranking and identifying extremes, with applications," in Bernardo, J.M., Degroot, M.H., Lindley, D.V., and Smith, A.F.M. (eds.), *Bayesian Statistics*, Oxford University Press, 5, 277-296.

Moura, F., and Holt, D. (1999), "Small area estimation using multilevel models," *Survey Methodology*, 25, 73-80.

Nandram, B. (1999), "An empirical Bayes prediction interval for the finite population mean of a small area," *Statistica Sinica*, 9, 325-344.

Nandram, B., and Sedransk, J. (1993), "Empirical Bayes estimation for the finite population mean on the current occasion," *Journal of the American Statistical Association*, 88, 994-1000.

Nepomnyaschy, L., and Reichman, N.E. (2006), "Low birthweight and asthma among young urban children," *American Journal of Public Health*, 96, 1604-1610.

Olsen, M. K., and Schafer, J. L. (2001), "A two-part random-effect model for semicontinuous longitudinal data," *Journal of the American Statistical Association*, 96, 730-745.

Otto, M.C., and Bell, W.R. (1995), "Sampling error modeling of poverty and income statistics for States," *Proceedings of the American Statistical Association*, Section on Government Statistics, 160-165.

Pfeffermann, D. (2002), "Small area estimation – New developments and directions," *International Statistical Review*, 70, 125-143.

Pfeffermann, D., and Tiller, R. (2005), "Bootstrap approximation to prediction MSE for state-space models with estimated parameters," *Journal of Times Series Analysis*, 26, 893-916.

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

Prasad, N.G.N., and Rao, J.N.K. (1999), "On robust small area estimation using a simple random effects model," *Survey Methodology*, 25, 67-72.

Prescott, P. (1978), "Selection of trimming proportions for robust adaptive trimmed means," *Journal of the American Statistical Association*, 73, 133-136.

Press, W.H., and Farrar, G.R. (1990), "Recursive stratified sampling for multidimensional monte carlo integration," *Computers in Physics*, 4, 190-195.

Rao, J.N.K. (2003), *Small Area Estimation,* New York: John Wiley and Sons.

Rao, J.N.K. (2005), "Interplay between sample survey theory and practice: an appraisal," *Survey Methodology*, 31, 117-138.

Rao, J.N.K., and Yu, M. (1994), "Small area estimation by combining time series and cross-sectional data," *Canadian Journal of Statistics*, 22, 511-528.

Raudenbush, S.W., Yang, M., and Yosef, M. (2000), "Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation," *Journal of Computational and Graphical Statistics*, 9, 141-157.

Robert, C.P., and Casella, G. (1999), *Monte Carlo Statistical Methods*, New York: Springer-Verlag.

Royall, R. M. (1970), "On finite population sampling under certain linear regression models," *Biometrika*, 57, 377-87.

Royall, R. M. (1976), "The linear least squares prediction approach to two-stage Sampling", *Journal of the American Statistical Association*, 71, 657-64.

Small Area Estimation Project (SAEP) Report (2003), "Model Based Small area Estimation Series," available at *http://www.statistics.gov.uk/methods_quality/downloads/small_area_est_report/.*

Sarndal, C.E., Swensson, B., and Wretman, J. (1992), *Model Assisted Survey Sampling*, New York: Springer-Verlag.

Scott, A., and Smith, T.M.F. (1969), "Estimation in multi-stage surveys," *Journal of the American Statistical Association*, 64, 830-840.

Singh, A.C., Folsom, R.E., Jr. and Vaish, A.K.. (2005), "Small area modeling for survey data with smoothed error covariance structure via generalized design effects," *Federal Committee on Statistical Methods Conference Proceedings*, Washington, D.C.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., and Van der Linde, A. (2002), "Bayesian measures of model complexity and fit (with discussion)," *Journal of the Royal Statistical Society*, *B*, 64, 583-640.

Steffensen, F.H., Sorensen, H.T., and Gillman, M.W. (2000), "Low birthweight and preterm delivery as risk factor for asthma and atopic dermatitis in young adult males," *Epidemiology*, 11, 185-188.

Stuart, A., Ord, K., and Arnold, S. (1999), *Kendall's Advanced Theory of Statistics,* 2A. London: Arnold, a member of the Hodder Headline Group.

Stroud, A. and Secrest, D. (1966), *Gaussian Quadrature Formulas*, Prentice Hall.

Stukel, D.M., and Rao, J.N.K. (1999), "Small area estimation under two-fold nested errors regression models," *Journal of Statistical Planning and Inference*, 78, 131-147.

Stroud, T. W. F. (1991), "Hierarchical Bayes predicative means and variances with application to sample survey inference," *Communications in Statistics - Theory and Methods*, 20, 13-36.

Tierney, L., and Kadane, J.B. (1986), "Accurate approximations for posterior moments and marginal densities," *Journal of the American Statistical Association*, 81, 82-86.

Tierney, L., Kass, R.E., and Kadane, J.B. (1989), "Fully exponential Laplace approximations to expectations and variances of nonpositive functions," *Journal of the American Statistical Association*, 84, 710-716.

Tomberlin, T. J. (1988), "Predicting accident frequencies for drivers classified by two factors," *Journal of the American Statistical Association***,** 83, 309-321.

Valliant, R. (1985), "Nonlinear prediction theory and the estimation of proportions in a finite population," *Journal of the American Statistical Association*, 80, 631-641.

Valliant, R (1986), "Mean squared error estimation of finite populations under nonlinear models," *Communications in statistics*, *A*, 15, 1975-1993.

Valliant, R. (1987), "Generalized variance functions in stratified two-stage sampling," *Journal of the American Statistical Association*, 82, 499-508.

Valliant, R., Dorfman, A.H., and Royall, R.M. (2000), *Finite Population Sampling and Inference: A Prediction Approach*, New York: Wiley and Sons.

Vijayan, K. (1968), "An exact π ps sampling scheme: generalization of a method of Hanurav," *Journal of the Royal Statistical Society, A*, 30, p. 556.

Wilcox, A.J. (2001), "On the importance — and the unimportance — of birthweight," *International Journal of Epidemiology*, 30, 1233-1241.

Wilcox, A.J., and Russell, I.T. (1983), "Birthweight and perinatal mortality: II. On weight-specific mortality," *International Journal of Epidemiology*, 12, 19-25.

Wilcox, A.J. (1993), "Birthweight and perinatal mortality: the effect of maternal smoking," *American Journal of Epidemiology*, 137, 1098-1104.

Wolfinger, R. (1993), "Laplace's approximation for nonlinear mixed models," *Biometrika*, 80, 791-795.

Wolter, K.M. (1985), *Introduction to Variance Estimation*, New York: Springer-Verlag.

Wong, G. Y., and Mason, W. M. (1985), "The hierarchical logistic regression model for multilevel analysis," *Journal of the American Statistical Association*, 80, 513-524.

Xie, D., Raghunathan, T.E., and Lepkowski, J.M. (2007), "Estimation of the proportion of overweight individuals in small areas – a robust extension of the Fay-Herriot model," *Statistics in Medicine*, 26, 2699-2715.

You, Y. (2008), "An integrated modeling approach to unemployment rate estimation for subprovincial areas of Canada," *Survey Methodology*, 34, 19-27.

You, Y., and Rao, J.N.K. (2002a), "Small area estimation using unmatched sampling and linking models," *Canadian Journal of Statistics*, 30, 3-15.

You, Y., and Rao, J.N.K. (2002b), "A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights," *Canadian Journal of Statistics*, 30, 431-439.

You, Y., Rao, J.N.K., and Gambino, J. (2003), "Model-based unemployment rate estimation for the Canadian Labour Force Survey: a hierarchical Bayes approach," *Survey Methodology*, 29, 25-32.