INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

**Bell & Howell Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600**

UMI[®]

Hierarchical Bayes and Related Methods for Model-Based Small Area Estimation

by
Yong You

A thesis submitted to
the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

School of Mathematics and Statistics

Carleton University
Ottawa, Ontario, Canada
September, 1999

© Copyright 1999

Yong You

National Library
of Canada

Acquisitions and
Bibliographic Services

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale
du Canada

Acquisitions et
services bibliographiques

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-48519-6

Canada

The undersigned hereby recommend to the Faculty of
Graduate Studies and Research acceptance of the thesis

**Hierarchical Bayes and Related Methods for Model-Based
Small Area Estimation**

Submitted by

Yong You, B.Sc., M.A.

in partial fulfillment of the requirements
for the degree of Doctor of Philosophy.

Kenneth S. Williams

Director, School of Mathematics and Statistics

J.N.K. Rao

Dr. J.N.K. Rao, Thesis Supervisor

J. N. K. PRAJAD

External Examiner

Carleton University

September, 1999

Abstract

This dissertation consists of nine chapters. Chapter 1 is an introduction to small area estimation. It presents a review of small area problems and related estimation approaches including empirical best linear unbiased prediction (EBLUP) approach and Bayes approach. EBLUP approach includes EBLUP estimators and mean squared error (MSE) approximation. Bayes approach includes empirical Bayes (EB) and hierarchical Bayes (HB) analysis. Chapter 2 gives a review of Markov Chain Monte Carlo (MCMC) sampling methods and computational hierarchical Bayes analysis methods. Chapters 3 to 8 contain several projects in the area of small area estimation. Chapter 9 gives some topics for future research.

In Chapter 3, a hierarchical Bayes analysis approach is applied to the Canadian Census undercoverage estimation problem. A robust hierarchical Bayes model is constructed for the direct survey Census undercoverage estimates and the posterior estimates of the undercoverage rate are obtained via Gibbs sampling method. Model selection and comparison are also provided.

In Chapter 4, hierarchical multi-level models including unequal error variance models and random error variance models are used to estimate small area means. Rao-Blackwellized estimators that reduce the Monte Carlo errors are also obtained. The estimation procedure is illustrated on a small data set from Brazil. Model selection and sensitivity analysis are also presented.

In Chapter 5, a pseudo hierarchical Bayes approach is presented to estimate small area means using small area models and sampling weights. We construct a pseudo posterior mean as an estimator for the small area mean and pseudo posterior

variance as a measure uncertainty of the estimator. A simulation study shows that the proposed method is very attractive and efficient. The pseudo Bayes approach is also extended to the general multi-level models considered in Chapter 4.

In Chapter 6, we extend the pseudo Bayes approach proposed in Chapter 5 to a nonlinear unit-level model in small area estimation. We also consider a nonlinear area-level model in this chapter and make a comparison of two approaches through a limited simulation study.

In Chapter 7, we consider binary data analysis in small area estimation. We provide a sampling based approach to estimate small area characteristics. A compound beta-binomial model is proposed to smooth direct survey estimates of response rates in a survey.

In Chapter 8, a cross-sectional and time series model is proposed for small area estimation. EBLUP estimators of small area means and their MSE approximations are obtained under this model. Hierarchical Bayes approach is also applied to this model. Rao-Blackwellized estimators of the posterior means and posterior variances of small area means are obtained.

Finally, some suggestions for further research are presented in Chapter 9.

Acknowledgements

I know this is a very special moment for me in my life. I have successfully finished my Ph.D. study. For so many years, I have been waiting for this moment. From China to Canada, from Chengdu to Toronto, from Toronto to Calgary, and from Calgary to Ottawa, my dream finally comes true. I would like to take this special opportunity to thank all the people who helped me make this dream come true during the past years.

First of all, I would like to express my sincere and deeply grateful thanks to my Ph.D. thesis supervisor, Professor J. N. K. Rao, for his supervision and encouragement during the course of this work, for the knowledge and skills he taught me in the areas of survey sampling and small area estimation, as well as for his kindness and support during the past four years.

I would like to thank my Ph.D. thesis committee members, Drs. D.R. Thomas and M. Mojirsheibani of Carleton University, Drs. M. Alvo and D. Krewski of University of Ottawa and Dr. N.G.N. Prasad of University of Alberta, for their time and hard work they put in for my thesis examination.

My thanks are also extended to Peter Dick and Diane Stukel of Statistics Canada for their helpful discussions and suggestions. I benefitted from Peter Dick in the area of Census undercoverage estimation. Diane Stukel provided good advice on my research projects such as binary data analysis.

I would like to thank Professors Claudia Czado and Amy Wu of York University in Toronto. Their support helped to make it possible for me to come to Carleton University in 1995. Also Professor Czado was my M.A. supervisor and I benefitted a

lot from her, especially in the area of Markov chain Monte Carlo sampling methods.

Finally, special thanks go to my wife Jiahua Li for her persistent support and love and to my parents in Chengdu, China for their consistent and great encouragement for so many years. In particular, I would like to take this opportunity to mention my father. Ke-Wei You, a professor in control theory and operation research, who always gives me encouragement and teaches me how to overcome various difficulties in my life.

This research was partially supported by Ontario Graduate Scholarship (1996-1999) and Assistantships from Carleton University.

Contents

Acceptance Sheet	ii
Abstract	iii
Acknowledgements	v
Contents	vii
List of Tables	xi
List of Figures	xiii
1 Introduction to Small Area Estimation	1
1.1 Introduction	1
1.2 Small Area Models	3
1.3 Model-based Estimation	5
1.3.1 EBLUP approach	5
1.3.2 Bayes approach	7
1.4 Thesis Outline	11

2	MCMC Sampling and Bayesian Analysis	13
2.1	Markov Chain Monte Carlo	13
2.2	Gibbs Sampling	15
2.2.1	Conditioning	15
2.2.2	The Gibbs sampler	16
2.2.3	Sampling issues	17
2.2.4	Computing tools	19
2.3	Metropolis-Hastings Algorithm	20
2.3.1	The Metropolis kernel	20
2.3.2	The General algorithm	21
2.3.3	Random walk chains	22
2.3.4	Independence Metropolis chains	23
2.3.5	Rejection Metropolis algorithm	24
2.3.6	Metropolized Gibbs sampling	26
2.4	Bayes Inference via MCMC	28
2.4.1	Hierarchical Bayes approach	28
2.4.2	Bayesian model choice	30
2.5	Summary	36
3	Census Undercoverage Estimation	38
3.1	Introduction	38
3.2	Models and Inferential Methodology	41
3.2.1	General model	41
3.2.2	Specification of fixed effects	44

3.2.3	Hierarchical Bayes modelling	45
3.2.4	Gibbs sampling inference	48
3.3	Census Undercoverage Estimation	49
3.3.1	Description of the survey estimates	49
3.3.2	Estimation and results	51
3.3.3	Bayesian model determination	53
3.4	Conclusion	57
4	Bayes Multi-level Models	65
4.1	Introduction	65
4.2	Hierarchical Bayes Multi-level Models	67
4.2.1	Equal error variance model	67
4.2.2	Unequal error variance model	72
4.2.3	Random error variance model	74
4.3	Data Analysis	76
4.3.1	Small area estimation	76
4.3.2	Model selection and comparison	81
4.3.3	Sensitivity analysis	84
4.4	Summary	85
5	Pseudo Bayes Small Area Estimation	95
5.1	Introduction	95
5.2	Simple Random Effects Model	97
5.3	Bayesian Analysis	99

5.3.1	Bayesian estimation	99
5.3.2	Pseudo Bayes estimation	101
5.3.3	A Gibbs sampling approach	103
5.4	A Simulation Study	105
5.5	Extension to Multi-level Models	110
6	Bayesian Analysis of Nonlinear Models	117
6.1	Introduction	117
6.2	A Unit-level Log Linear Model	119
6.2.1	Predictor of random effect	119
6.2.2	Pseudo Bayes estimator	120
6.2.3	Monte Carlo evaluation	122
6.2.4	Remarks	124
6.3	An Area-level Model	126
6.3.1	A transformation approach	127
6.3.2	A complete sampling approach	129
6.4	A Numerical Study	131
7	Small Area Binary Data Analysis	137
7.1	Introduction	137
7.2	Small Area Estimation	138
7.2.1	Hierarchical probit model	138
7.2.2	Inferential methodology	140
7.3	Smoothing Small Area Estimates	143

7.3.1	Background	143
7.3.2	Hierarchical beta-binomial model	145
7.3.3	Gibbs sampling inference	147
7.3.4	Data analysis	149
7.3.5	Remarks	151
8	Small Area Time Series Data Analysis	157
8.1	Introduction	157
8.2	Cross-sectional and Time Series Model	158
8.3	BLUP Estimator	160
8.4	EBLUP Estimator	161
8.4.1	Estimation of σ_{ε}^2	161
8.4.2	Estimation of σ_v^2	163
8.4.3	EBLUP estimator of θ_{iT}	164
8.5	MSE Estimation	164
8.5.1	Second-order approximation to the MSE	164
8.5.2	Second-order approximation to estimator of MSE	167
8.6	Hierarchical Bayes Approach	168
8.6.1	The hierarchical Bayes model	168
8.6.2	Gibbs sampling inference	169
9	Future Research	171
	Bibliography	174

List of Tables

3.1	Model specification	58
3.2	Direct and Posterior Provincial Undercoverage Estimates	59
3.3	Provincial Efficiency Comparison	60
3.4	Bayesian Cross-Validation Predictive Residuals	61
3.5	CPO and log Likelihood Values	62
3.6	Pointwise CPO Comparison	63
4.1	Summary of Small Area Data	87
4.2	Estimated Posterior Small Area Means	88
4.3	Simulation Standard Errors	89
4.4	Estimated Posterior Variance of μ_i	90
4.5	Estimated Sampling Error Variances	91
4.6	Log of the Product of the CPO Values	91
4.7	Comparison of Estimated Small Area Means	92
4.8	Comparison of Estimated Sampling Error Variances	93
5.1	Comparison of $\hat{\theta}_i^p$ and $\hat{\theta}_i^m$ Under Population 1	114
5.2	Comparison of $\hat{\theta}_i^p$ and $\hat{\theta}_i^m$ Under Population 2	114
5.3	Comparison of $\hat{\theta}_i^p$ and $\hat{\theta}_i^m$ Under Population 3	115

5.4	Relative difference of $\hat{\theta}_i^p$ and $\tilde{\theta}_i^c$	115
5.5	Results of $\hat{V}^n(\theta_i)/\hat{V}^p(\theta_i)$	116
6.1	Relative Differences of Posterior Means	135
6.2	Ratio of Posterior Variances	136
7.1	Results of RHGs in Newfoundland	155
7.2	Results of RHGs in P.E.I.	156
7.3	Results for All Provinces	156

List of Figures

3.1	CPO Plot of Census Undercoverage Models	64
4.1	Multi-level Models CPO Plot	94
7.1	Estimates and Standard Errors (Newfoundland)	152
7.2	Estimates and Standard Errors (P.E.I.)	153
7.3	Provincial Level Response Rate Estimates	154

Chapter 1

Introduction to Small Area Estimation

1.1 Introduction

Sample surveys, for most purposes, have taken the place of complete enumeration or census as a more effective means of obtaining information on wide-ranging topics of interest at frequent intervals over time. Sample surveys are generally designed to provide reliable direct estimates for large areas and major domains (subgroups) of a population. For example, the U.S. National Health Interview Survey provides reliable estimates of health characteristics, such as mortality, disability and utilization of health services, for the nation and four broad geographical regions. The Canadian Labour Force Survey produces reliable monthly estimates of unemployment rate for the nation and all provinces. Increasingly, such surveys are also being used to provide estimates for small areas, which are needed in formulating policies and programs, in allocation of government funds, in regional programs, and so on. The term “small area” is commonly used to denote a small geographical area, such as a county, a municipality or a census division. It may also describe a small domain,

i.e., a small subpopulation such as a specific age-sex-race group of people within a large geographical area. However, the usual direct survey estimates for a small area, based on data only from the sample units in the area, are likely to yield unacceptably large standard errors due to unduly small size of the sample in the area. Sample sizes for small areas are typically small because the overall sample size in a survey is usually determined to provide specific accuracy at a much higher level of aggregation than that of small areas. This makes it necessary to “borrow strength” from related areas to find more accurate estimates for a given area or, simultaneously, for several areas. Such estimates are based on either implicit or explicit models that provide a link to related small areas through supplementary data such as administrative records and recent census counts. These small area estimators may be classified into two groups: (1) indirect estimators based on implicit models, and (2) model-based estimators. Indirect estimators proposed in the literature include synthetic, sample-size-dependent, and composite, while group (2) model-based estimators include empirical best linear unbiased prediction (EBLUP), empirical Bayes (EB), and hierarchical Bayes (HB). Ghosh and Rao (1994) give a comprehensive review of the indirect estimation methods for small areas.

In recent years, model-based estimators have been widely used in small area estimation due to the development of the EBLUP and EB theories and Markov chain Monte Carlo (MCMC) sampling methods for Bayes analysis. For the EBLUP method, the major milestone is the work on the mean squared error (MSE) approximation of EBLUP estimators (Kackar and Harville, 1984; Prasad and Rao, 1990). After the important work of Prasad and Rao (1990), several EBLUP estimators and

their MSE estimators have been proposed under various small area models including time series model (Rao and Yu, 1994), two-fold nested-error regression model (Stukel and Rao, 1999), random variance mixed linear model (Kleffe and Rao, 1992), robust small area model (Lahiri and Rao, 1995) and multi-level models (Holt and Moura, 1999). For the Bayes analysis, after the fundamental paper on MCMC sampling by Gelfand and Smith (1990), computational Bayes methods and hierarchical Bayes methods have been applied to various aspects of statistical inference, including small area estimation (Datta and Ghosh, 1991; Farrell, 1998; Ghosh et al., 1998; Malec et al., 1997). It is now generally accepted that one should use indirect estimators based on explicit models that relate the small areas through supplementary data such as administrative records and recent census counts. An advantage of the model approach is that it permits validation of models from the sample data. In the rest of this section, we review some basic models used in the small area estimation as well as the inferential methodologies including EBLUP approach and Bayes approach.

1.2 Small Area Models

Small area models may be broadly classified into two types. The first type is area-specific or area-level model, and the second type is element-specific or unit-level model. For the area-level model, only area-specific auxiliary data $x_i = (x_{i1}, \dots, x_{ip})^T$ are available for the areas $i = 1, \dots, m$. The parameter of interest for the i th small area, for example, the small area mean μ_i , is assumed to be related to x_i through a linear model with random area effect v_i :

$$\mu_i = x_i^T \beta + v_i, \quad i = 1, \dots, m, \quad (1.1)$$

where β is the vector of regression parameters and the v_i 's are uncorrelated with mean zero and variance σ_v^2 ; normality of the v_i 's are often assumed. For making inference about the μ_i 's, we assume that direct survey estimates, $\hat{\mu}_i$, are available and it is customary to assume that

$$\hat{\mu}_i = \mu_i + e_i, \quad i = 1, \dots, m, \quad (1.2)$$

where e_i 's are sampling errors and are independent $N(0, \psi_i)$ with known sampling variance ψ_i . Combining the sampling model (1.2) with the linking model (1.1), we get the area-level random effect model as:

$$\hat{\mu}_i = x_i^T \beta + v_i + e_i, \quad i = 1, \dots, m. \quad (1.3)$$

Note that model (1.3) involves both design-based random variable e_i and model-based random variable v_i . If the sampling variances ψ_i are unknown, smoothing of estimated variances $\hat{\psi}_i$ is often done to get stable estimates ψ_i^* which are treated as the known true ψ_i . The area-level model (1.3) is also referred as the well-known Fay-Herriot model (Fay and Herriot, 1979).

For the second type of models, i.e., the unit-level model, element-specific auxiliary data $x_{ij} = (x_{ij1}, \dots, x_{ijp})^T$ are available for the population units. Unit values y_{ij} are assumed to be related to x_{ij} through a nested error regression model:

$$y_{ij} = x_{ij}^T \beta + v_i + e_{ij}, \quad j = 1, \dots, N_i, i = 1, \dots, m, \quad (1.4)$$

where $v_i \stackrel{iid}{\sim} N(0, \sigma_v^2)$ are independent of $e_{ij} \stackrel{iid}{\sim} N(0, \sigma_e^2)$, and N_i is the number of population units in the i th area. The mean for the i th area may be written as

$$\mu_i = \bar{x}_i^T \beta + v_i, \quad (1.5)$$

assuming that N_i is large, where $\bar{x}_i$ is the vector of known means of the x_{ij} for the i th area.

Note that both the area-level model (1.3) and the unit-level model (1.4) are special cases of the following general mixed linear model with block diagonal covariance structure:

$$Y = X\beta + Zu + e, \quad (1.6)$$

where Y is the vector of sample observations, X and Z are known matrices, and u and e independent random vectors with $u \sim (0, G)$ and $e \sim (0, R)$. The covariance matrices G and R may depend on some parameters θ called variance components.

The small area mean μ_i is a linear combination of fixed effects β and realized value of random effect u_i . Thus it is a special case of $\mu = l_1^T \beta + l_2^T u$.

1.3 Model-based Estimation

We now present the EBLUP, EB and HB approaches to small area estimation in the context of models (1.3), (1.5), and (1.6).

1.3.1 EBLUP approach

For the general mixed linear model (1.6), it follows from Henderson (1975) that for θ known, the BLUP estimator of $\mu = l_1^T \beta + l_2^T u$ is given by

$$\bar{\mu}(\theta, Y) = l_1^T \bar{\beta} + l_2^T GZ^T V^{-1}(Y - X\bar{\beta}) \quad (1.7)$$

where $\bar{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} Y$ is the generalized least squares estimator of β and $V = R + ZGZ^T$ is the variance-covariance matrix of Y . $\bar{\mu}$ depends on the variance components θ , but in practice θ will be unknown. Replacing θ by a suitable

consistent estimator $\hat{\theta} = \hat{\theta}(Y)$, we obtain a two-stage estimator $\hat{\mu} = \bar{\mu}(\hat{\theta}, Y)$ which is called empirical BLUP or EBLUP estimator.

A measure of variability associated with the estimator $\hat{\mu}$ is given by the mean squared error, $MSE(\hat{\mu}) = E(\hat{\mu} - \mu)^2$. Kackar and Harville (1984) showed that

$$MSE(\hat{\mu}) = MSE(\bar{\mu}) + E(\hat{\mu} - \bar{\mu})^2 \quad (1.8)$$

under normality of v and e , provided that $\hat{\theta}$ is translation-invariant, i.e., $\hat{\theta}$ remains unchanged when y is changed to $-y$ or to $y - Xa$ for all y and a . For the first term, $MSE(\bar{\mu})$, exact expression can be found as:

$$MSE(\bar{\mu}) = g_1(\theta) + g_2(\theta), \quad (1.9)$$

where

$$g_1(\theta) = l_2^T (G - GZ^T V^{-1} ZG) l_2, \quad (1.10)$$

and

$$g_2(\theta) = (l_1 - X^T V^{-1} ZG^T l_2)^T (X^T V^{-1} X)^{-1} (l_1 - X^T V^{-1} ZG^T l_2). \quad (1.11)$$

For the second term $E(\hat{\mu} - \bar{\mu})^2$, following Kacker and Harville's Taylor series approximation (1984), Prasad and Rao (1990) proposed a further approximation, given by

$$E(\hat{\mu} - \bar{\mu})^2 \approx \text{tr} \left[\left(\frac{\partial b^T}{\partial \theta} \right) \cdot V \cdot \left(\frac{\partial b^T}{\partial \theta} \right)^T \cdot E(\hat{\theta} - \theta)(\hat{\theta} - \theta)^T \right] \equiv g_3(\theta), \quad (1.12)$$

where $b^T = l_2^T GZ^T V^{-1}$. Thus combining (1.9) and (1.12), Prasad and Rao (1990) obtained second-order approximation to $MSE(\hat{\mu})$ as:

$$MSE(\hat{\mu}) = g_1(\theta) + g_2(\theta) + g_3(\theta) \quad (1.13)$$

which is correct to term of order $O(m^{-1})$ as the number of small areas, m , increases. They also provided an estimator of the MSE as:

$$mse(\hat{\mu}) = g_1(\hat{\theta}) + g_2(\hat{\theta}) + 2g_3(\hat{\theta}), \quad (1.14)$$

which is correct to $O(m^{-1})$ in the sense $E(mse(\hat{\mu})) = MSE(\hat{\mu}) + o(m^{-1})$. Using these general theoretic results, Prasad and Rao (1990) obtained the EBLUP estimators and their corresponding MSE estimators under the nested error regression model, random regression coefficient model and Fay-Herriot model. Stukel and Rao (1999) employed a two-fold nested error regression model for small areas and obtained EBLUP estimators of small area means under this model and the corresponding MSE estimators. Rao and Yu (1994) extended the Fay-Herriot model to time series data and considered a cross sectional and time series model involving a stationary autoregressive model for the time component. Kleffe and Rao (1992) considered a simple random effect linear model with random error variances and obtained the EBLUP estimators of small area means and the corresponding MSE approximations. For the Fay-Herriot model, Lahiri and Rao (1995) established the robustness property of the MSE estimator to non-normality of the small area random effect v_i in the sense that it remains valid under non-normality.

1.3.2 Bayes approach

In Bayes analysis, the fundamental tool is the Bayes theorem and hierarchical modelling. Two approaches are commonly used in statistical data analysis, namely, empirical Bayes (EB) approach and hierarchical Bayes (HB) approach. In the EB approach, the posterior distribution of the parameters of interest given the data is

first obtained, assuming that the model hyperparameters are known. The hyperparameters are estimated from the marginal distribution of the data, and inferences are then based on the estimated posterior distribution. Whereas in the HB approach, prior distributions on the hyperparameters are specified and the posterior distribution of the parameters of interest is then obtained. Inferences are based on the posterior distribution; in particular, a parameter of interest is estimated by its posterior mean and its precision is measured by its posterior variance.

Typically a hierarchical Bayes model involves $(f(x|\theta), \pi(\theta|\lambda))$, where θ is the parameter of interest and λ is the hyperparameter. Then the posterior distribution of θ given λ is

$$f(\theta|x, \lambda) = \frac{f(x|\theta)\pi(\theta|\lambda)}{\int f(x|\theta)\pi(\theta|\lambda)d\theta}. \quad (1.15)$$

By using (1.15), we can find the posterior mean $E(\theta|x, \lambda)$ and posterior variance $V(\theta|x, \lambda)$ of θ . However, the posterior mean and posterior variance depend on the hyperparameter λ , which is usually unknown in practice. Then the HB approach introduces an additional prior distribution on the hyperparameter λ and finally the posterior mean and posterior variance of θ can be calculated as

$$E(\theta|x) = \int E(\theta|x, \lambda)d\lambda, \quad (1.16)$$

and

$$V(\theta|x) = E_\lambda[V(\theta|x, \lambda)] + V_\lambda[E(\theta|x, \lambda)], \quad (1.17)$$

where E_λ and V_λ respectively denote the expectation and variance with respect to the posterior distribution of λ given x . Usually noninformative or approximately noninformative prior distribution is used for the hyperparameter λ .

The EB approach, however, proposes to estimate the hyperparameter λ from the marginal distribution

$$m(x|\lambda) = \int f(x|\theta)\pi(\theta|\lambda)d\theta \quad (1.18)$$

by $\hat{\lambda}(x)$ and to use $f(\theta|x, \hat{\lambda}(x))$ as a naive posterior distribution. $\hat{\lambda}(x)$ can be derived from the marginal distribution $m(x|\lambda)$ by using various frequentist methods such as the method of moments, the maximum likelihood method, or yet another method. Thus a naive EB approach uses $E(\theta|x, \hat{\lambda}(x))$ as the estimator of θ and measures its uncertainty by the naive posterior variance $V(\theta|x, \hat{\lambda}(x))$. This can lead to severe underestimation of the true posterior variance $V(\theta|x)$, although $E(\theta|x, \hat{\lambda}(x))$ is approximately equal to the true posterior mean $E(\theta|x)$ because it follows from (1.17) that $V(\theta|x, \hat{\lambda}(x))$ is a good approximation only to the first variance term on the right side of (1.17) and the second variance term is ignored in the naive EB approach, that is, it fails to take account of the uncertainty about the hyperparameter λ . Two methods of accounting for the underestimation of the true posterior variance have been proposed in the literature. The first method is based on the bootstrap (Laird and Louis, 1987), while the second method uses an asymptotic approximation to the posterior variance $V(\theta|x, \hat{\lambda}(x))$ irrespective of the form of the prior on λ . Morris (1983) gave an excellent account of the EB approach and significant applications. Ghosh and Rao (1994) gave an excellent review on EB approach in small area estimation. For the EB in small area applications, see, for example, Datta et al. (1997), Arora et al. (1997) and Farrell (1998).

Compared to the EB approach, the HB approach is straightforward and clear-cut but computationally intensive, often involving high dimensional integration. Recent

advances in computational aspects of the HB approach, such as Gibbs sampling method (Gelfand and Smith, 1990), however, seem to overcome the computational difficulties to a large extent. The HB approach looks promising, but caution should be exercised in applying the HB approach. For example, improper prior on a hyperparameter may lead to improper posteriors. Hobert and Casella (1996) demonstrated that the Gibbs sampler could lead to seemingly reasonable inferences about a nonexistent posterior distribution. This happens when the posterior is improper and yet the full conditional distributions for the Gibbs sampler are all proper and well-defined. One way to avoid improper posteriors is to use proper priors as suggested in Hobert and Casella (1996). In mixed random effects models, ignorance can be modeled by using a normal prior with very large variance for the fixed effects and inverse gamma priors with very small parameter values for the variance components. This strategy is also used in the well-known Gibbs sampling software BUGS (Spiegelhalter et al., 1996).

Ghosh et al. (1998) applied the HB approach to generalized linear models with random effects. Employing diffuse gamma proper priors on the inverse of the variance components, they provide conditions for the posterior to be proper. These conditions require the scale parameter of the gamma distribution to be strictly greater than zero. But as the scale parameter approaches zero to reflect lack of prior information, we are essentially approximating an improper posterior by a proper posterior. For other HB applications in small area estimation, we refer to Ghosh et al. (1996), Datta et al. (1998) and Malec et al. (1997).

1.4 Thesis Outline

In this section, we present an outline of this dissertation. Chapter 2 gives a comprehensive review of Markov chain Monte Carlo (MCMC) sampling methods and related topics on computational Bayesian analysis. From Chapter 3 to Chapter 7, we apply the hierarchical Bayes approach to various problems under different small area models. In particular, in Chapter 3, we extend the Fay-Herriot model to a robust hierarchical model in the context of Canadian Census undercoverage estimation and obtain Bayesian estimators of the provincial census undercoverage rates. In Chapter 4, we extend the general multi-level model by allowing unequal error variances or by assuming a random effect model for the error variances. We fit our models to a small area data set from Brazil and make appropriate Bayesian inference and model selections. In Chapter 5, we propose a pseudo hierarchical Bayes approach to estimate small area means by combining small area models and sampling weights. We also extend the proposed pseudo hierarchical Bayes approach to multi-level models. In Chapter 6, we consider nonlinear models in small area estimation. Two types of models are considered, i.e., unit-level and area-level models. We extend the pseudo Bayes approach proposed in Chapter 5 to a nonlinear unit-level model for small area estimation. We also consider a nonlinear area-level model in this chapter and make a comparison of two approaches through a limited simulation study. And in Chapter 7, we consider binary data analysis in small area estimation. We first present a sampling-based approach to estimate small area means in the binary data case and then we present a compound beta-binomial model to smooth the direct estimates of small area response rates with application to a real survey of Statistics

Canada. In Chapter 8, we consider a cross-sectional and time series model. EBLUP estimators and MSE approximations are obtained under this model. Hierarchical Bayes approach is also applied to this model. And finally, Chapter 9 presents some possible future research directions and application areas.

Chapter 2

MCMC Sampling and Bayesian Analysis

2.1 Markov Chain Monte Carlo

A time homogeneous Markov chain with invariant (or stationary) distribution π is a sequence of random variables $\{X_n : n \geq 0\}$ such that the transition kernel $p(x, y)$ where $\int p(x, y)dy = 1$ satisfies

$$\pi(y) = \int p(x, y)\pi(x)dx$$

and the n th iterate is given by $p^{(n)}(x, y) = \int p^{(n-1)}(x, z)p(z, y)dz$, where $p^{(1)}(x, y) = p(x, y)$. The crucial result is that if a chain has a proper invariant distribution π and it is irreducible and aperiodic, then π is the unique invariant distribution and is also the equilibrium distribution of the chain, i.e., π is the limiting distribution.

The basic idea of Markov chain Monte Carlo (MCMC) methods is to create a Markov chain $\{X_0, X_1, X_2, \dots\}$ with invariant distribution π . Here the invariant distribution π is known (probably up to a normalizing constant). It is the target density from which samples are desired, but the transition kernel is unknown. To

generate samples from $\pi(\cdot)$, the MCMC methods find and utilize a transition kernel $p(x, y)$ whose n th iterate converges to $\pi(\cdot)$ for large n . The process is started at an arbitrary X_0 and iterated a large number of times. After a sufficiently long “burn-in” of say m iterations, the observations $\{X_t : t = m + 1, \dots, n\}$ generated from the simulation can be regarded as dependent samples approximately from the target density $\pi(\cdot)$. Inference can be based on the output from the Markov chain. For instance, to evaluate $E[f(X)]$ with X having density $\pi(\cdot)$, we can simply use sample mean

$$\bar{f} = \frac{1}{n - m} \sum_{t=m+1}^n f(X_t).$$

This is called an ergodic average. Convergence to the required expectation is ensured by the ergodic theorem.

To construct such a Markov chain, we need to find an appropriate transition kernel $p(x, y)$. A $p(x, y)$ that satisfies the following “reversibility condition”

$$\pi(x)p(x, y) = \pi(y)p(y, x) \tag{2.1}$$

has $\pi(\cdot)$ as its invariant distribution (Chib and Greenberg, 1995). This condition is a sufficient condition to be satisfied by $p(x, y)$. Many approaches are available for constructing Markov chains with a specified invariant distribution. Gibbs sampling and Metropolis-Hastings algorithm are the most two important Monte Carlo methods available based on differently specified transition kernels. The following sections will give details of these two MCMC methods. A careful theoretical discussion of MCMC methods and related applications can be found in Smith and Roberts (1993), Tierney (1994) and Brooks (1998).

2.2 Gibbs Sampling

Gibbs sampling is a MCMC method to simulate complex multivariate non-normalized distributions and to obtain marginal distributions. For many high dimensional non-normalized distributions, Gibbs sampling is an iterative MCMC method that enables us to avoid the curse of dimensionality by sampling low dimensional densities and to obtain inference about the joint and marginal distributions. The most prominent application is for inference within a Bayesian framework. In the Bayesian context, inference are based on posterior distributions of model parameters. Samples produced by the Gibbs sampler can be used straightforwardly for Bayesian inference.

2.2.1 Conditioning

Gibbs sampling uses method of conditioning and the so called “product of kernels principle” (Chib and Greenberg, 1995) to construct Markov chains with invariant distributions. Suppose that the target density is $\pi(x)$, and now suppose that x can be partitioned into two or more sets of parameters called blocks, i.e., $x = (x_1, \dots, x_r)$, $r \geq 2$. For illustration, let $x = (x_1, x_2)$. The full conditional distributions are defined as $[x_1|x_2]$ and $[x_2|x_1]$ with densities $\pi_{1|2}(x_1|x_2)$ and $\pi_{2|1}(x_2|x_1)$, respectively. It can be shown that the transition kernel given by

$$p((x_1, x_2), (y_1, y_2)) = \pi_{1|2}(y_1|x_2)\pi_{2|1}(y_2|y_1)$$

has $\pi(\cdot)$ as its invariant distribution. Moreover, under rather general conditions, iteration of $p(x, y)$ converges to $\pi(x)$ (Roberts and Smith, 1994; Chan, 1993; Tierney, 1994). The sampler proceeds as follows: choose an arbitrary starting value for x_2

and generate a value of y_1 . Then generate a value of y_2 given the value of y_1 . This completes one iteration. Now generate a new value of the first block, conditional on y_2 , and continue these iterations for a large number of runs. The Gibbs sampler thus transforms a joint distribution involving several blocks into a series of conditional distributions involving just one block. This is the strategy used in the Gibbs sampler of Gelfand and Smith (1990).

2.2.2 The Gibbs sampler

The Gibbs sampler was introduced as a MCMC tool in the context of image restoration by Geman and Geman (1984). Gelfand and Smith (1990) offered the Gibbs sampler as a very general approach for fitting statistical models. They extended the applicability of the work of Geman and Geman and also broadened the substitution sampling ideas which Tanner and Wong (1987) proposed under the name of data augmentation. Gibbs sampling has become a widely employed computational tool for carrying out statistical inference in complex models.

Suppose we have a set of random variables $X = (x_1, \dots, x_k)$, where x_i 's are viewed as blocks. These blocks might be univariate but generally they could be multivariate as well. Then given an arbitrary set of starting values $X^{(0)} = (x_1^{(0)}, \dots, x_k^{(0)})$, the Gibbs sampler is a Markovian updating scheme that proceeds as follows: if the current state of X is $X^{(t)} = (x_1^{(t)}, \dots, x_k^{(t)})$, then we make the transition to $X^{(t+1)}$ by sampling:

$$\begin{aligned} x_1^{(t+1)} & \text{ from } h(x_1 | x_2^{(t)}, x_3^{(t)}, \dots, x_k^{(t)}) \\ x_2^{(t+1)} & \text{ from } h(x_2 | x_1^{(t+1)}, x_3^{(t)}, \dots, x_k^{(t)}) \\ & \vdots \\ x_k^{(t+1)} & \text{ from } h(x_k | x_1^{(t+1)}, x_2^{(t+1)}, \dots, x_{k-1}^{(t+1)}) \end{aligned}$$

The $h(x_i|x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_k)$'s are referred as the full, or complete, conditional distributions and the process of updating each of the k blocks as indicated updates the entire vector X producing one complete iteration of the Gibbs sampler. After n such iterations, we obtain $(x_1^{(n)}, \dots, x_k^{(n)})$. Under mild conditions, the following results hold (Geman and Geman, 1984; Gelfand and Smith, 1990):

1. $(x_1^{(n)}, \dots, x_k^{(n)}) \xrightarrow{d} [x_1, \dots, x_k]$ and hence for each x_i , $x_i^{(n)} \xrightarrow{d} [x_i]$ as $n \rightarrow \infty$.
2. The joint distribution of $(x_1^{(n)}, \dots, x_k^{(n)})$ converges at an exponential rate to $[x_1, \dots, x_k]$ as $n \rightarrow \infty$.
3. For any measurable function T of $x_1, \dots, x_k$ whose expectation exists,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{l=1}^n T(x_1^{(l)}, \dots, x_k^{(l)}) \xrightarrow{a.s} E(T(x_1, \dots, x_k))$$

Hence for large n , $(x_1^{(n)}, \dots, x_k^{(n)})$ can be regarded as an observation from the joint distribution of $X = (x_1, \dots, x_k)$. The samples can be used directly for statistical inference.

2.2.3 Sampling issues

Conceptually, the Gibbs sampler emerges as a rather straightforward algorithmic procedure. For implementation of the Gibbs sampler, we need to draw samples from the full conditional distributions $h(x_i|x_j, j \neq i)$. There are basically two possibilities:

- (1) $h(x_i|x_j, j \neq i)$ is a standard distribution. Particularly in the Bayesian framework, $h(x_i|x_j, j \neq i)$ generally could be a (multivariate) normal distribution or a (multivariate) inverse gamma distribution. Sampling from these standard distributions is straightforward.

(2) $h(x_i|x_j, j \neq i)$ does not have a closed form. In Bayesian framework, $h(x_i|x_j, j \neq i)$ could be a posterior density proportional to a non-normalized likelihood function. For this kind of situation, the following methods are considered:

- Envelope rejection method as discussed in Dagpunar (1988).
- Adaptive rejection sampling for univariate log concave densities developed by Gilks and Wild (1992).
- Hybrid algorithms with Metropolis-Hastings algorithm. These algorithms typically include Metropolis within Gibbs, rejection Metropolis sampling and adaptive rejection Metropolis sampling within Gibbs. See Gilks et al (1995) and Chib and Greenberg (1995) for general discussions.

For sampling from full conditional distributions, a related issue is the introduction of latent variables. Often full conditional distributions can be simplified by augmenting the model with suitable latent variables. Though overall dimensionality is increased, much easier sampling of the full conditional distributions results since non-standard conditional distributions become standard ones. Illustrative examples can be found in, e.g., Albert and Chib (1993).

Another consideration in the area of iterative simulation is whether to run a single long run or to run several parallel chains. The Gibbs sampler can be used to generate G separate chains, each of length n , and the final states from each chain are retained. The debate's result is not clear (Gelman and Rubin, 1992). Experience and theoretical assessments in the simulation literature appear to favor the use of a single or a small number of long runs (Tierney, 1994). It is now generally agreed that

running many short chains, motivated by a desire to obtain independent samples from the target density, is misguided unless there is some special reason for needing independent samples. The major concern here is the convergence and the effect of starting values. Also for long runs, it is possible to use all the data after possibly discarding a small initial fraction and to get more information from the simulated chains.

Many other important issues about Gibbs sampling like convergence diagnostics, starting values, choice of parametrization, optimal inference, etc, can be found, for example, in Brooks, 1998; Tierney, 1994, etc.

2.2.4 Computing tools

Gibbs sampling has emerged as a dominant sampling-based approach for carrying out statistical inference in complex models, especially in the Bayesian framework. BUGS (Bayesian Inference Using Gibbs Sampling) is a computer program that carries out Bayesian inference on statistical problems using Gibbs sampling (Spiegelhalter et al, 1995). BUGS assumes a Bayesian or full probability model, in which all quantities are treated as random variables. The model consists of many unknown quantities but for which substantial conditional independence assumptions are appropriate. Particular structures include generalised linear models with hierarchical or crossed random effects, latent variable models, measurement errors in responses and covariates, informative censoring, constrained estimation, and missing data; see BUGS manual for examples of typical applications. Another suite of software, CODA (Convergence Diagnosis and Output Analysis) (Best et al, 1995), is a menu-driven set of S-Plus functions which serves as an output processor for the

BUGS. CODA may be used to perform convergence diagnostics and statistical and graphical output analysis of the simulated values from BUGS.

There are currently restrictions on the classes of model that can be fitted using BUGS. BUGS can handle many standard Bayes hierarchical models mentioned above, but not for many other complex and non-standard models (see BUGS manual for details). More efficient algorithms including Metropolis-Hastings algorithm and many combining strategies should be considered in research and practice.

2.3 Metropolis-Hastings Algorithm

The Metropolis-Hastings (M-H) algorithm introduced by Metropolis et al. (1953) and extended by Hastings (1970) is a different Markov chain method than the Gibbs sampler to simulate complex distributions. The M-H algorithm has been used extensively in physics, but despite the paper by Hastings it was little known to statisticians until recently. Papers by Tierney (1994) and Chib and Greenberg (1995) gave a review and statistical applications of the M-H algorithm.

2.3.1 The Metropolis kernel

The basic idea of the M-H algorithm is to use a candidate generating density to generate samples to construct a Markov chain with a given limiting distribution through certain acceptance-rejection steps. Suppose we have a candidate generating density denoted by $q(x, y)$, where $\int q(x, y)dy = 1$. This density is to be interpreted as saying that when a process is at the point x , the density generates a value y from $q(x, y)$. To construct a Markov chain with invariant distribution $\pi(\cdot)$, we define a

Metropolis kernel as

$$p(x, y) = \begin{cases} q(x, y)\alpha(x, y) & \text{if } x \neq y \\ 0 & \text{if } x = y \end{cases}$$

where $\alpha(x, y)$ is determined such that the kernel $p(x, y)$ satisfies the reversibility condition given by (2.1). $\alpha(x, y)$ is called the Metropolis probability of move and is defined as

$$\alpha(x, y) = \begin{cases} \min\left\{\frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}, 1\right\} & \text{if } \pi(x)q(x, y) > 0 \\ 1 & \text{if } \pi(x)q(x, y) = 0 \end{cases} \quad (2.2)$$

Under mild regularity conditions, the Metropolis kernel has $\pi(\cdot)$ as its invariant density. The regularity conditions required are irreducibility and aperiodicity. What these mean is that, if x and y are in the domain of $\pi(\cdot)$, it must be possible to move from x to y in a finite number of iterations with nonzero probability and the number of moves required to move from x to y is not required to be a multiple of some integers. These conditions are usually satisfied if $q(x, y)$ has a positive density on the same support as that of $\pi(\cdot)$. See Tierney (1994) and Chib and Greenberg (1995) for details.

2.3.2 The General algorithm

Suppose that we wish to sample from a target density $\pi(x)$, where x could be a random vector. Let $q(x, y)$ be a candidate generating density, where $\int q(x, y)dy = 1$. Define the probability of move $\alpha(x, y)$ as in (2.2), then the Metropolis-Hastings algorithm proceeds as follows:

1. Choose an arbitrary initial point $x^{(0)}$ and set $t = 0$.
2. Generate a candidate y from $q(x^{(t)}, \cdot)$ and U from $U(0, 1)$.

3. Let $x^{(t+1)} = y$ if $U \leq \alpha(x^{(t)}, y)$; otherwise let $x^{(t+1)} = x^{(t)}$.
4. Let $t = t + 1$ and go to step 2.

Several points about this algorithm should be noted. (1) The M-H algorithm is specified by its candidate generating density $q(x, y)$. (2) If a candidate value is rejected, the current value is taken as the next value in the sequence. (3) The calculation of $\alpha(x, y)$ does not require knowledge of the normalizing constant of $\pi(\cdot)$, since it appears both in the numerator and denominator. (4) If $\pi(y)q(y, x) \geq \pi(x)q(x, y)$, the chain moves to y , otherwise it moves with probability given by $\pi(y)q(y, x)/\pi(x)q(x, y)$.

The Metropolis-Hastings algorithm is very general, allowing a variety of useful special cases based on different choices of the candidate generating density $q(x, y)$. In order to implement the M-H algorithm, it is necessary that a suitable candidate generating density be specified. Considerable recent work is being devoted to the question of how these choices should be made. Tierney (1994) provides a systematic taxonomy of the kinds of choices available. Related issues can be found, for example, in Chib and Greenberg (1995), Gilks et al. (1995), etc.

2.3.3 Random walk chains

One family of candidate generating densities is given by setting $q(x, y) = f(y - x)$, where $f(\cdot)$ is a multivariate density function. The candidate y thus is generated according to the process $y = x + z$, where z is called the increment random variable and has density $f(\cdot)$. Since the candidate y is equal to the current value x plus an independent draw z from $f(\cdot)$, this Metropolis chain is called a random walk chain.

Possible choices of $f(\cdot)$ include multivariate normal distributions and multivariate t distributions. When $f(\cdot)$ is symmetric about the origin, i.e., $f(z) = f(-z)$, then $q(x, y) = q(y, x)$, the probability of move simplifies to

$$\alpha(x, y) = \min\left\{\frac{\pi(y)}{\pi(x)}, 1\right\}$$

which is the original Metropolis probability of move in Metropolis et al. (1953).

2.3.4 Independence Metropolis chains

The candidate generating density can also be chosen as a fixed density f . This option is discussed in Hastings (1970). In this case, $q(x, y) = f(y)$ and the probability of move $\alpha(x, y)$ can be written as

$$\alpha(x, y) = \min\left\{\frac{w(y)}{w(x)}, 1\right\}$$

where $w(x) = \pi(x)/f(x)$ is called the importance weight (or importance ratio). This sampling scheme is called “Metropolized independent sampling” and is closely related to the corresponding importance sampling process. Liu (1995) provides a comparison of this sampling scheme to rejection and importance sampling schemes. For the choice of f , we can choose f to be a multivariate normal or multivariate t density. For the independence sampling scheme to work well, f should be a good approximation to π , but it is safe if f is heavier-tailed than π . In some situations, a multivariate normal density might be tried, with mean at the mode of $\pi(\cdot)$ and covariance matrix somewhat greater than the inverse Hessian matrix

$$\left[-\frac{d^2 \log \pi(x)}{dx^T dx}\right]^{-1}$$

evaluated at the mode. The best choice of f is still open to question (Tierney, 1994; Chib and Greenberg, 1995).

If the target density $\pi(\cdot)$ can be written as $\pi(x) \propto \psi(x)h(x)$, where $h(x)$ is a density that can be sampled and $\psi(x)$ is uniformly bounded, then we can use $h(\cdot)$ as the candidate generating density and set $q(x, y) = h(y)$. In this case, the probability of move requires only the computation of the $\psi(\cdot)$ function (not π or h) and is given by

$$\alpha(x, y) = \min\left\{\frac{\psi(y)}{\psi(x)}, 1\right\}.$$

2.3.5 Rejection Metropolis algorithm

An interesting and important special case of the Metropolized independent sampling occurs if sampling from f uses envelope rejection sampling.

Envelope rejection sampling

In contrast to the MCMC methods, envelope rejection sampling is a method to generate iid samples from a target density $\pi(\cdot)$. Suppose $\pi(x) = K \cdot h(x)$, where $h(x)$ is a known unnormalized density, and K is the (possibly unknown) normalizing constant. Assume that $h(x) \leq cf(x)$ for all x , where $f(x)$ is a known density (envelope function) that can be simulated exactly and c is a known constant. Then to generate a variate from $\pi(x)$, the following rejection sampling procedure can be used:

- (i) Generate a candidate x^* from $f(\cdot)$ and a value R from $U(0, 1)$;
- (ii) If $R \leq \frac{h(x^*)}{cf(x^*)}$ then return $x = x^*$, otherwise reject x^* and go to step (i).

It is easily shown that the accepted variate x has pdf $\pi(x)$ (Dagpunar, 1988). The probability of acceptance (the sampling efficiency) is c^{-1} . In practice, however, it is very difficult to choose a dominating constant c such that c is large enough for $cf(x)$ to dominate $h(x)$ without choosing c excessively large, leading to an inefficient algorithm with many rejections. The rejection step can be optimized by setting

$$c = \sup_x \frac{h(x)}{f(x)};$$

even this choice, however, may result in an undesirably large number of rejections.

Rejection sampling chains

In the envelope rejection method, a constant c and a density $f(x)$ are needed such that $cf(x)$ dominates the unnormalized density $h(x)$ for all x with $\pi(x) = Kh(x)$ being the target density. Finding a c that does the trick may be difficult in some applications; moreover, choosing c excessively large can lead to unacceptably high proportion of rejected candidates. Fortunately, using envelope rejection method to generate an independence Metropolis chain provides a simple remedy. Suppose $c > 0$ is a known constant, but that $cf(x)$ does not necessarily dominate $h(x)$ for all x . We define a set $A = \{x : h(x) \leq cf(x)\}$ where domination occurs. Because $cf(x)$ does not dominate $h(x)$ on A^c , it follows that the target density $\pi(x)$ is not adequately sampled. This can be corrected with an Metropolis-Hastings rejection step applied to the values that come through the envelope rejection step. Define the Metropolis acceptance probability $\alpha(x, y)$ as

$$\alpha(x, y) = \begin{cases} 1 & \text{if } x \in A \\ \frac{cf(x)}{h(x)} & \text{if } x \notin A, y \in A \\ \min\{\frac{h(y)f(x)}{h(x)f(y)}, 1\} & \text{if } x \notin A, y \notin A \end{cases} \quad (2.3)$$

then the algorithm proceeds as follows: if the chain is currently at $x^{(n)} = x$, generate a candidate y independent of the current value x by applying the envelope rejection method with $cf(\cdot)$ as the “dominating” density. With probability $\alpha(x, y)$ given by (2.3), this candidate is accepted and the chain moves to $x^{(n+1)} = y$. Otherwise y is rejected and the $x^{(n+1)} = x$. Thus the algorithm will occasionally reject candidates when the chain is at a point $x \notin A$. This repeats the point x within the sample path, and thus compensates for the deficiency in the envelope at x . Dependence is introduced to make up for this deficiency and ensure that $\pi(\cdot)$ is the invariant distribution of the Markov chain. If the envelope is in fact adequate, i.e., $h(x) \leq cf(x)$ for all x , then no candidates are rejected, and the chain produces an iid sequence from π . However, when the bound $cf(\cdot)$ is inadequate, the rejection steps occasionally happen.

2.3.6 Metropolized Gibbs sampling

The Gibbs sampler given in section 2.2 requires that the complete full conditionals $h(x_i|x_j, j \neq i)$ be available for sampling. It is common in practice that $h(x_i|x_j, j \neq i)$ is not easy to sample from, especially when $h(x_i|x_j, j \neq i)$ does not have a closed form. One might use the envelope rejection method to get around it. It would require finding a constant c such that an appropriate candidate generating density dominates $h(x_i|x_j, j \neq i)$ which makes it impractical to use in the context of the Gibbs sampler because each draw involves a different conditional distribution with possibly a different constant c . For problems where an upper bound for the constant c over all possible conditionals can be at least approximately given such algorithms have been suggested and used for example in Zeger and Karim (1991) and Czado

and You (1993).

Alternatively, one can first approximate $h(x_i|x_j, j \neq i)$ by a candidate generating density $p(x_i)$ that is easy to sample from, such as normal or t distributions, and then correct the discrepancy by inserting a Metropolis-Hastings rejection step. This scheme can be used very generally by using different candidate generating densities in the Metropolis procedure within the Gibbs sampler. Moreover, it is not necessary to stop the Gibbs sampler to iterate the M-H algorithm; one can generate one value from the Metropolis-Hastings rejection step and continue with the next Gibbs step. For illustration, consider x with two blocks, i.e., $x = (x_1, x_2)$, and suppose that $h(x_1|x_2)$ is a standard density and easy to sample from, and $h(x_2|x_1)$ does not have a closed form. For Gibbs sampling, assume $x^{(n)}$ is the current state of the Markov chain. To generate $x^{(n+1)}$:

- (i) Generate $x_1^{(n+1)}$ from $h(x_1|x_2^{(n)})$.
- (ii) Generate a candidate y from a candidate generating density $p(x_2)$, with probability $\alpha(x_2^{(n)}, y)$ accept y , i.e., $x_2^{(n+1)} = y$, otherwise $x_2^{(n+1)} = x_2^{(n)}$. $\alpha(x, y)$ is the Metropolis acceptance probability given by (2.2) in general form.

Thus the algorithm is a hybrid of the Gibbs sampler and the M-H algorithm. Just one iteration of Metropolis rejection step suffices to preserve the stationary distribution of the Gibbs chain. However, this Markov chain may be slower to converge through the Metropolis rejection steps. For the choice of the candidate generating density, it seems to be a common wisdom that the candidate generating density $p(x_i)$ should have heavier tails than the true conditional density $h(x_i|x_j, j \neq i)$. For more

details on the hybrid scheme, see for example Gelman and Rubin (1993), Tierney (1994) and Brooks (1998).

2.4 Bayes Inference via MCMC

2.4.1 Hierarchical Bayes approach

We begin by specifying the overall hierarchical Bayes model in general terms. A hierarchical Bayes (HB) model normally consists of three parts. The sampling (or measurement) model specifies the distributions of the data given the parameters. The population (or structural) model expresses the relationships among the parameters. The parameters may include quantities of substantive interest that are not directly observable, as well as latent variables that are not of intrinsic interest but that facilitate specification of the model. Finally, the hyperparameter model specifies the prior distribution of the hyperparameters, which in general are variance and covariance components and regression parameters of the population model.

Bayesian inferences are statements about the posterior distribution of the parameters of interest. Inferences are based on the posterior distribution; in particular, a parameter of interest is estimated by its posterior mean and its precision is measured by its posterior variance. In the HB model, we are interested in obtaining the posterior estimates of parameters of interest given the data. The HB approach is straightforward and clear-cut but computationally intensive, often involving high dimensional integration. By applying the MCMC sampling methods, such as Gibbs sampling, however, we can overcome the computational difficulties to a large extent.

In Bayesian inference one is interested in sampling from the posterior distribution

of the parameters. For this assume that $y_i|\theta$ has conditional density $f_i(y|\theta)$ for $1 \leq i \leq n$ and that the prior information about $\theta = (\theta_1, \dots, \theta_k)$ is summarized by a prior density $\pi(\theta)$. Let $\pi(\theta|Y)$ denote the posterior density of θ given the data $Y = (y_1, \dots, y_n)$. Since

$$\pi(\theta|Y) = \frac{\prod_{i=1}^n f_i(y_i|\theta)\pi(\theta)}{\int \prod_{i=1}^n f_i(y_i|\theta)\pi(\theta)d\theta}$$

it might be difficult to sample from $\pi(\theta|Y)$ directly due to the high dimensional integration with respect to θ . However, one can construct a Markov chain $\{\theta^t = (\theta_1^t, \dots, \theta_k^t), t \geq 0\}$ with limiting distribution $\pi(\theta|Y)$ using the Gibbs sampler. For this we generate $\theta^{(t)} = (\theta_1^{(t)}, \dots, \theta_k^{(t)})$, $t = 1, \dots, m$ by sampling

$$\begin{aligned} \theta_1^{(t)} &\text{ from } \pi(\theta_1|\theta_2^{(t-1)}, \theta_3^{(t-1)}, \dots, \theta_k^{(t-1)}, Y) \\ \theta_2^{(t)} &\text{ from } \pi(\theta_2|\theta_1^{(t)}, \theta_3^{(t-1)}, \dots, \theta_k^{(t-1)}, Y) \\ &\vdots \\ \theta_k^{(t)} &\text{ from } \pi(\theta_k|\theta_1^{(t)}, \theta_2^{(t)}, \dots, \theta_{k-1}^{(t-1)}, Y). \end{aligned}$$

Under regularity conditions, the distribution of $\theta^{(t)} = (\theta_1^{(t)}, \dots, \theta_k^{(t)})$ converges to $\pi(\theta|Y) = \pi(\theta_1, \dots, \theta_k|Y)$ as $t \rightarrow \infty$. The marginal inference about $\pi(\theta_i|Y)$ can be based on the marginal sample $\theta_i^{(t)}$, $t = 1, \dots, m$. Gelfand and Smith (1990) noticed that the full conditional structure of the sampling algorithm and the samples $\{\theta_j^{(t)}, j \neq i\}$, $t = 1, \dots, m$ should be exploited. Indeed, if the quantity of interest is $E(\theta_i|Y)$, one can use the average of the conditional expectations

$$\tilde{\theta}_i = \frac{1}{m} \sum_{t=1}^m E[\theta_i|\theta_j^{(t)}, j \neq i, Y] \quad (2.4)$$

instead of the empirical sample mean

$$\hat{\theta}_i = \frac{1}{m} \sum_{t=1}^m \theta_i^{(t)}. \quad (2.5)$$

This modification is based on the Rao-Blackwell theorem and $\tilde{\theta}_i$ is called the Rao-Blackwellized estimate of the posterior mean of θ_i . $\hat{\theta}_i$ is seen to be dominated by $\tilde{\theta}_i$ (Gelfand and Smith, 1990; Liu, Wong and Kong, 1994).

To estimate the posterior variance of θ_i , by noting that

$$\text{Var}(\theta_i|Y) = E[\text{Var}(\theta_i|\theta_j, j \neq i, Y)] + \text{Var}[E(\theta_i|\theta_j, j \neq i, Y)], \quad (2.6)$$

we obtain the following mixture estimate for the posterior variance of θ_i

$$\begin{aligned} \tilde{V}(\theta_i) &= \frac{1}{m} \sum_{t=1}^m \text{Var}(\theta_i|\theta_j^{(t)}, j \neq i, Y) + \frac{1}{m} \sum_{t=1}^m [E(\theta_i|\theta_j^{(t)}, j \neq i, Y)]^2 \\ &\quad - \frac{1}{m^2} [\sum_{t=1}^m E(\theta_i|\theta_j^{(t)}, j \neq i, Y)]^2 \end{aligned} \quad (2.7)$$

We can also obtain the marginal density estimate for θ_i by using the average of the full conditional densities

$$\frac{1}{m} \sum_{t=1}^m \pi(\theta_i|\theta_j^{(t)}, j \neq i, Y) \quad (2.8)$$

instead of using regular kernel estimation methods (Gelfand and Smith, 1990). It should be noted that the full conditional distribution $\pi(\theta_i|\theta_j, Y)$ must have a closed form if one wants to use the Rao-Blackwellized estimates for θ_i ; otherwise, $E(\theta_i|\theta_j^{(t)}, j \neq i, Y)$ and $V(\theta_i|\theta_j^{(t)}, j \neq i, Y)$ are not explicitly available.

2.4.2 Bayesian model choice

MCMC methods enable the fitting of Bayesian models of virtually unlimited complexity for the analysis of complicated statistical problems. In extremely complicated hierarchical and random effects models, with the generality of MCMC methods there is temptation to fit models so large that their parameters are unidentified, or nearly

so, which finally requires insightful reparametrizations of the model. In practice, responsible data analysis should address the issue of model determination, which usually includes (i) is a given model adequate? and (ii) within a collection of models under consideration, which is the best? In this section we will briefly discuss how to use Bayesian predictive distributions via MCMC sampling methods for model determination.

Predictive distributions

A model is defined by the specification of a joint distribution for observables and unobservables. Suppose Y is a set of observations $\{y_i; i = 1, \dots, n\}$ and θ is the model unobservables. A *prior predictive* density is defined as the marginal density of Y

$$f(Y) = \int f(Y|\theta)f(\theta)d\theta \quad (2.9)$$

where $f(\theta)$ is the prior distribution of θ . That is, the conditional density of the observables is averaged over prior knowledge about the unobservables. This suggests that many other predictive densities might be examined.

If Y_{obs} denotes the actual observations, the *posterior predictive* density, $f(Y|Y_{obs})$, is the predictive density of a new independent set of observables, Y , under the model, given the actual observables. It would be computed as

$$f(Y|Y_{obs}) = \int f(Y|\theta) \cdot f(\theta|Y_{obs})d\theta \quad (2.10)$$

since $f(Y|\theta) = f(Y|\theta, Y_{obs})$. The posterior predictive density averages the conditional density of the observables, against the posterior knowledge about the unobservables. The posterior predictive density offers a conservative view of model

performance. If the observations $Y = Y_{obs}$ are not in agreement with $f(Y|Y_{obs})$, which is derived under the model using these observations, then the model is unlikely to receive support under any other criterion. By marginalizing $f(Y|Y_{obs})$, we obtain

$$f(y_i|Y_{obs}) = \int f(y_i|\theta) \cdot f(\theta|Y_{obs})d\theta \quad (2.11)$$

We can use $f(y_i|Y_{obs})$ to generate samples and construct confidence intervals for y_i , which can be compared with $y_{i,obs}$.

Gelfand et al. (1992) create diagnostics based on the *cross-validation predictive densities* $\{f(y_i|Y_{(i)}); i = 1, \dots, n\}$, where $Y_{(i)}$ is the data vector with i th point y_i deleted. The density $f(y_i|Y_{(i)})$ is attractive in that it suggests what values of y_i are likely when the model is fitted to all of the observations except y_i . The actual $y_{i,obs}$ can then be compared with this density in a variety of ways to see whether it is likely under the model.

Cross-validation predictive densities are usually calculated by the equation:

$$f(y_i|Y_{(i)}) = \int f(y_i|\theta, Y_{(i)}) \cdot f(\theta|Y_{(i)})d\theta \quad (2.12)$$

rather than through the equivalent equation, $f(y_i|Y_{(i)}) = f(Y)/f(Y_{(i)})$. In the case of conditionally independent observations given θ , $f(y_i|\theta, Y_{(i)}) = f(y_i|\theta)$, then $f(y_i|Y_{(i),obs})$ and $f(y_i|Y_{obs})$ should be quite similar unless $y_{i,obs}$ is very influential in the fittings.

Model determination

In Bayesian model determination, traditionally the *Bayes factor* plays an important role. If Y_{obs} denotes the actual observations and $f(Y|M_i)$ denotes the marginal

density under model $M_i, i = 1, 2$, then the Bayes factor is defined as

$$BF = \frac{f(Y_{obs}|M_1)}{f(Y_{obs}|M_2)} \quad (2.13)$$

which provides the relative weight of evidence for model M_1 compared to model M_2 . The Bayes factor arises formally as the ratio of the posterior odds for M_1 versus M_2 to the prior odds for M_1 versus M_2 .

One important use of the Bayes factor is as a summary of the evidence for model M_1 against model M_2 provided by the data. It can be useful to consider twice the logarithm of the Bayes factor, which is on the same scale as the familiar deviance and likelihood-ratio test statistics. Rough calibration of the Bayes factor has been proposed by Jeffreys (1961) and Raftery (1995).

The product of cross-validation predictive densities $\prod_{i=1}^n f(y_i|Y_{(i),obs})$ has been proposed as a surrogate for $f(Y)$, giving rise to the *pseudo-Bayes factor*:

$$\prod_{i=1}^n \frac{f(y_{i,obs}|Y_{(i),obs}, M_1)}{f(y_{i,obs}|Y_{(i),obs}, M_2)} \quad (2.14)$$

as a surrogate for the Bayes factor (Gelfand et al., 1992).

For model choice, Pettit and Young (1990) proposed the *conditional predictive ordinate* (CPO) as

$$CPO_i = f(y_{i,obs}|Y_{(i),obs}). \quad (2.15)$$

Small CPO_i indicates that $y_{i,obs}$ does not support the model. One of the principal attractions of CPO_i arises from the obvious relationship

$$\log CPO_i = \log f(Y_{obs}) - \log f(Y_{(i),obs}).$$

Thus the $\log(CPO)$ for the i th observation contrasts the prior predictive density of all the observations with that of all but the i th observation, and may be interpreted

as the incremental contribution to the adequacy of the model attributable to the i th observation. Consider the CPO ratio:

$$C_i = \frac{f(y_{i,obs}|Y_{(i),obs}, M_1)}{f(y_{i,obs}|Y_{(i),obs}, M_2)}$$

and note that $\prod_{i=1}^n C_i$ is just the previously defined pseudo-Bayes factor and

$$\log C_i = \log BF - \log BF_{(i)},$$

where $BF_{(i)}$ is the Bayes factor for model M_1 relative to model M_2 using all but the i th observation. If $C_i > 1$, i.e., $\log C_i > 0$, $y_{i,obs}$ supports model M_1 ; if $C_i < 1$, i.e., $\log C_i < 0$, it supports M_2 . We can plot $\log C_i$ versus i to see how many observations support each model, which observations support which model and the strength of the support.

In choosing among more than two models, we can no longer work with CPO ratios. However, we can still plot CPO_i versus i for each model to compare the models visually. With such a CPO plot we can see which models do better than others, which points are poorly fit under all models, which models are essentially indistinguishable, and so on.

Computational issues

In general, output from a MCMC simulation can be utilized to estimate the predictive densities and implement the model determination. Assume that the model has been fitted by a MCMC method yielding a collection θ_j^* ; $j = 1, 2, \dots, m$, which we treat as a sample from $f(\theta|Y_{obs})$. Then for predictive density estimation, we have:

1. Prior predictive density estimation

To calculate the prior predictive density $f(Y)$, for any proper density $h(\theta)$ note that

$$(f(Y))^{-1} = \int \frac{h(\theta)}{f(Y|\theta)f(\theta)} \cdot f(\theta|Y) d\theta,$$

then a Monte Carlo integration yields

$$\hat{f}(Y) = \frac{1}{\frac{1}{m} \sum_{j=1}^m \frac{h(\theta_j^*)}{f(Y|\theta_j^*)f(\theta_j^*)}}. \quad (2.16)$$

Here $h(\theta)$ plays the role of an importance sampling density for the posterior, so the better h approximates the posterior the better $\hat{f}(Y)$ will be. If $f(\theta)$ is proper, we might take $h(\theta) = f(\theta)$, then (2.16) becomes the harmonic mean of the $f(Y|\theta_j^*)$. However, since $f(\theta)$ and $f(\theta|Y)$ will usually be quite different, typically this estimate is very unstable.

2. Posterior predictive density estimation

For $f(Y|Y_{obs})$, an immediate Monte Carlo integration from (2.10) gives its density estimation as

$$\hat{f}(Y|Y_{obs}) = \frac{1}{m} \sum_{j=1}^m f(Y|\theta_j^*). \quad (2.17)$$

Thus marginally we have

$$\hat{f}(y_i|Y_{obs}) = \frac{1}{m} \sum_{j=1}^m f(y_i|\theta_j^*).$$

3. Cross-validation predictive density estimation

For the cross-validation predictive density, $f(y_i|Y_{(i)})$ can be written as

$$\begin{aligned} f(y_i|Y_{(i)}) &= \frac{f(Y)}{f(Y_{(i)})} \\ &= \frac{1}{\int \frac{f(Y_{(i)}, \theta)}{f(Y, \theta)} \cdot f(\theta|Y) d\theta} \end{aligned}$$

$$= \frac{1}{\int \frac{1}{f(y_i|Y_{(i),\theta})} \cdot f(\theta|Y) d\theta}.$$

Thus an immediate Monte Carlo integration yields

$$\hat{f}(y_i|Y_{(i),obs}) = \frac{1}{\frac{1}{m} \sum_{j=1}^m \frac{1}{f(y_i|Y_{(i),obs}, \theta_j^*)}}, \quad (2.18)$$

the harmonic mean of the $\{f(y_i|Y_{(i),obs}, \theta_j^*): j = 1, \dots, m\}$. If the $\{y_i\}$ are conditionally independent given θ , $f(y_i|Y_{(i),obs}, \theta_j^*)$ simplifies to $f(y_i|\theta_j^*)$. Then the estimate of the conditional predictive ordinate (CPO) is given by

$$\widehat{CPO}_i = \hat{f}(y_{i,obs}|Y_{(i),obs}) = \frac{1}{\frac{1}{m} \sum_{j=1}^m \frac{1}{f(y_{i,obs}|\theta_j^*)}} \quad (2.19)$$

4. Sampling from predictive densities

Suppose we wish to sample from $f(Y|Y_{obs})$. If for each θ_j^* , we sample a data set Y_j^* from $f(Y|\theta_j^*)$, then marginally Y_j^* is a sample from $f(Y|Y_{obs})$. Consequently $y_{i,j}^*$, the i th element of Y_j^* , is a sample from $f(y_i|Y_{obs})$. Usually, this straightforward sampling suffices to investigate model adequacy. For instance, suppose we generate a sample $\{y_{i1}^*, \dots, y_{im}^*\}$ from the posterior predictive density $f(y_i|Y_{obs})$, where m need not be too large, say 100 to 200. We could note how consistent $y_{i,obs}$ is with these y_{ij}^* by using these samples to create $(1 - \alpha)$ equal tail predictive intervals for each i . We would then expect roughly $(1 - \alpha)$ of the $y_{i,obs}$ to fall in their respective intervals.

2.5 Summary

In this Chapter, Markov Chain Monte Carlo (MCMC) sampling methods are reviewed. MCMC methods including Gibbs sampling and Metropolis algorithm have

been found to be very powerful techniques for practical hierarchical Bayes analysis in many statistical areas, including model-based approaches small area estimation in survey sampling. In survey sampling, traditional methods, including demographic techniques and design-based methods, can be combined with model-based methods to obtain more accurate estimates. Typically various fixed and random effects models are used for small area estimation in survey sampling to “borrow strength” from related areas to find more accurate estimates for a given area or, simultaneously, for several areas. Considerable attention has been given to such model-based methods in recent years. See Ghosh and Rao (1994) for a review.

In the hierarchical Bayes (HB) approach, prior distributions on the model parameters are specified and the posterior mean is obtained as the estimate of the parameters. Posterior variance is used as a measure of uncertainty associated with the HB estimator (posterior mean). Gibbs sampling method is used in complex models to overcome computational difficulties in Bayesian analysis. However, caution should be exercised in using hierarchical Bayes models and Gibbs sampling. Problems including Bayesian model checking and improper posteriors are the most important ones. The Gibbs sampler may lead to proper conditional posteriors even when the joint posterior is improper (Hobert and Casella, 1996).

Finally, the MCMC method and Bayesian analysis go together naturally. MCMC is a simple idea with enormous potential.

Chapter 3

Hierarchical Bayes Estimation of Census Undercoverage

3.1 Introduction

The Census of Canada is conducted every five years. The census, however, does not enumerate all the inhabitants that should fill a census form on Census Day. In the 1991 Canadian census, it is estimated that about 3% of the population were not enumerated. Thus the census needs to be adjusted for undercoverage in order to properly represent the demographic picture of the country on Census Day. Statistics Canada estimates the undercoverage through two sample surveys. The gross undercoverage was measured using the Reverse Record Check (RRC). The Overcoverage Study estimates persons double counted or erroneously included in the census count. The direct net undercoverage estimate was obtained by subtracting the overcoverage estimate from the RRC undercoverage estimate.

In 1991, in a major departure from the established procedure, it was decided to revise the population estimates to agree with the Census counts adjusted to account for the estimated difference between gross undercoverage and gross overcoverage, or

net undercoverage, in the Census. The new base population was formed by adding the net provincial undercoverage estimate to the provincial Census count. This created an adjusted base upon which all the other population figures were derived using modelling and demographic methods.

Leading up to this decision, the focus of the research was on whether the provincial Census count adjusted for net undercoverage in the Census was an improvement on the provincial Census count alone. The result of this research was the development of a procedure similar to a preliminary test (Royce, 1992). From this test and other consideration, Statistics Canada decided that the Census adjusted for the net undercoverage estimated from the coverage studies was the best estimate on which to base the population estimates program.

There is an extensive literature and a lot of controversy surrounding the estimation of census undercoverage and its use in the population estimates. Prior to the 1991 Census, the population estimates were based solely on the Census results. In Census years, the annual population estimates were essentially the Census counts (with some minor adjustments for the Census reference date). Between Censuses, the provincial population estimates are obtained by taking into account the births, deaths, immigrants, emigrants and returning Canadians during the intercensal period. When the next Census results became available, the population estimates were revised to agree with the new Census counts. In 1991, the population estimates were based on the Census counts adjusted for the estimated net undercoverage in the Census. The technical criteria for deciding in favour of adjustment was a procedure known as the preliminary test. This test uses the results of the coverage

studies to decide between a full adjustment and no adjustment. The results of this procedure indicated that the Census counts with the net undercoverage added in at the provincial level were an improvement on the Census counts alone with regard to both the estimates of the provincial populations and the estimates of the provincial shares of the national population.

After the release of the 1991 coverage studies' results, a debate was started on examining various estimators of the provincial undercoverage. The selection of an estimator for the undercoverage is based on statistical criteria - weighted mean square error. Rivest(1995) proposed a composite estimator that used the national undercoverage rate as a synthetic estimate. The effect of this composite estimator was to shrink all provincial undercoverage rates towards the national rate. Provinces close to the national rate moved relatively little while provinces far from the national rate moved by a more substantial amount. Rivest's estimator is a compromise between the direct survey estimator and a synthetic estimator. However, Rivest's composite estimator performs poorly for Prince Edward Island (PEI) and Ontario, which have respectively the smallest and the largest undercoverage rate. For these two provinces, the Rivest composite estimator is less efficient than the direct survey estimator (You, 1996).

In recent years considerable attention has been given to applying modelling techniques to improve the direct undercoverage estimates from sample surveys. Datta et al. (1992) and Zaslavsky (1993) proposed hierarchical and empirical Bayes methods for the adjustment of census undercoverage based on US census count and situations. Dick (1995) proposed a regression model utilizing an empirical Bayes method

to estimate the undercoverage in small domains. The objective of this Chapter is to present a hierarchical Bayes model for Canadian census undercoverage. This is a continuing work of my internship project with Statistics Canada in 1996 (You, 1996). The provincial undercoverage rate is estimated by its posterior mean and its precision is measured by its posterior variance. A Gibbs sampling method is used to simulate samples from the posterior distributions.

Section 2 presents a general hierarchical Bayes model and specification of fixed and random effects for Census undercoverage estimation. A Gibbs sampling inference procedure is also illustrated to address the undercoverage estimation problem. Rao-Blackwellized estimators of the posterior mean and posterior variance are obtained. Section 3 presents the results of using the hierarchical Bayes models for 1991 undercoverage estimation. Bayesian model determination including investigation of model adequacy and model choice is also discussed. And section 4 presents some conclusions and possible future research directions.

3.2 Models and Inferential Methodology

3.2.1 General model

Suppose there are n provinces and in the i th province the Census has counted C_i persons while an unknown number U_i persons were missed by the Census. The coverage studies provide an estimate $\hat{U}_i$ of the net undercoverage along with an associated *known* variance ξ_i^2 . One objective of the Population Estimates Program is to estimate the true population of the i th province on Census Day.

The true population of the i th province can be written as $T_i = C_i + U_i$. Since the

Census count is observed without sampling error, most of the work in constructing a model centres around the estimate of missed persons. The coverage studies use a sample survey which through standard estimation procedures produces an estimate of the number of missed persons. The model generally used to describe this estimation situation is written as

$$\hat{U}_i = U_i + \epsilon_i, \quad i = 1, \dots, n,$$

where $\epsilon_i \sim N(0, \xi_i^2)$ with known ξ_i^2 . This model assumes that the estimates $\hat{U}_i$'s are design unbiased, normally distributed with known sampling variances. These assumptions may be restrictive, in particular the estimates of missed persons $\hat{U}_i$ are certainly subject to some unknown bias. The normal distribution assumption of the sampling errors at the provincial level seems quite reasonable due to the Central Limit Theorem. The sampling variances ξ_i^2 are estimated through generalized variance function models of the form $V(\hat{U}_i) \propto C_i^\gamma$ (Dick, 1995) and treated as known. Thus ξ_i^2 is actually a smoothed estimate of the sampling variance.

Interpreting the differences in missed persons between provinces is difficult because of the large differences in provincial sizes in Canada. A more meaningful measure is the undercoverage rate which is defined as

$$\theta_i = \frac{U_i}{U_i + C_i} = \frac{U_i}{T_i},$$

where θ_i is the true undercoverage rate for the i th province. The direct estimate of θ_i can be obtained as

$$y_i = \frac{\hat{U}_i}{\hat{U}_i + C_i} = \frac{\hat{U}_i}{\hat{T}_i}.$$

Based on this transformation, we assume that the sampling model for θ_i can be written as

$$y_i = \theta_i + e_i, \quad (3.1)$$

where $e_i \sim N(0, \psi_i^2)$. The sampling variance ψ_i^2 is known and is calculated from a Taylor approximation as

$$\psi_i^2 = \text{Var}(y_i) = \frac{(1 - \theta_i)^4}{C_i^2} \text{Var}(\hat{U}_i) \approx \frac{(1 - y_i)^4}{C_i^2} \xi_i^2.$$

The true undercoverage rate θ_i is assumed to be related to a set of provincial auxiliary variables $x_i = (x_{i1}, \dots, x_{ip})^T$ through a regression model given by

$$\theta_i = x_i^T \beta + v_i, \quad i = 1, \dots, n, \quad (3.2)$$

where β is the regression parameters and v_i is provincial random effect with $E(v_i) = 0$ and $\text{Var}(v_i) = \sigma^2$, where σ^2 is unknown.

Combining the sampling model and the regression model, we obtain the model:

$$y_i = x_i^T \beta + v_i + e_i, \quad i = 1, \dots, n, \quad (3.3)$$

which involves design-induced random variables, e_i , and model based random effects v_i . The model (3.3) is a special case of the general mixed linear models discussed in Chapter 1. Ericken and Kadane (1985) used the same kind of models for census undercount estimation. Zaslavsky (1993) also used a linear mixed models to estimate population shares. However, we note that the above modelling has the following limitations: (1) The sampling variance ψ_i^2 is approximated by $\frac{(1-y_i)^4}{C_i^2} \xi_i^2$ and then is treated as known. That is a very strong assumption, since ψ_i^2 is indeed a function of the unknown mean θ_i ; (2) In (3.2), assuming v_i to be normal may not be appropriate

because θ_i is the undercoverage rate which is between 0 and 1. A more realistic model may use $\log(\theta_i) = x_i^T \beta + v_i$. But this will lead to a non-linear mixed model, which is more complicated but it can be handled by using the method developed in Section 6.3, Chapter 6.

3.2.2 Specification of fixed effects

The fixed effect part of the general model relates the true undercoverage rate to underlying set of auxiliary variables. The simplest model sets the regression component equal to some fixed value. A logical choice would be the national undercoverage rate given by

$$r = \frac{\sum_{i=1}^n U_i}{\sum_{i=1}^n T_i}.$$

Hence, we would equate the regression component to the national undercoverage rate as $x_i^T \beta = r$. This model can be made more flexible by the following approach.

For any province, we assume that the fixed effect can be written as $x_i^T \beta = \tau + \alpha_i$, where τ is a fixed value and α_i is a pre-specified value related to the i th province. Since $E(\theta_i) = \tau + \alpha_i$, the national undercoverage rate may be written approximately in terms of τ and α_i as

$$r = \frac{\sum_{i=1}^n \theta_i T_i}{\sum_{i=1}^n T_i} \approx \frac{\sum_{i=1}^n (\tau + \alpha_i) T_i}{\sum_{i=1}^n T_i},$$

which leads to

$$\tau = r - \sum_{i=1}^n \alpha_i p_i,$$

where $p_i = T_i / \sum T_k$ is the share of the total population of the i th province and is

assumed known. Hence the undercoverage rate for any province can be written as

$$\theta_i = r + \alpha_i - \sum_{k=1}^n \alpha_k p_k + v_i. \quad (3.4)$$

This expression has the following properties: First, note that if $\alpha_i = 0$ for all $i = 1, \dots, n$, then $\theta_i = r + v_i$. Secondly, for any non-zero value of α_i , the national undercoverage total is preserved. This can be seen by multiplying (3.4) by the provincial population total T_i and summing this quantity up for the n provinces and taking expectation to show that both sides total $\sum U_k$.

The expression (3.4) allows the pre-specified α_i to be written in terms of relative differences in undercoverage rate between provinces. For instance, for any two provinces, we have

$$E(\theta_i - \theta_j) = \alpha_i - \alpha_j.$$

This implies that α_i 's represent the difference in undercoverage rate between provinces. Thus by pre-specifying how the Census undercoverage rate is expected to differ between provinces we can specify the fixed components without regard for national undercoverage rate r .

By combining the sampling model (3.1) and model (3.4) for θ_i , we obtain a joint model for the direct survey undercoverage estimate y_i as follows:

$$y_i = r + \alpha_i - \sum_{k=1}^n \alpha_k p_k + v_i + e_i, \quad i = 1, \dots, n. \quad (3.5)$$

3.2.3 Hierarchical Bayes modelling

Having determined the form of the fixed effects part of the model, the random effects will now be introduced. The basic framework of the hierarchical Bayes model

follows from Lindley and Smith (1972) discussion of exchangeability. The first stage assumes that the undercoverage rates for each province are unbiasedly estimated with known sampling variance. The second stage assumes that the true undercoverage rates for each province are unbiasedly estimated by the national undercoverage rate adjusted for differences between the provinces with an unknown variance. The errors associated with this stage are assumed to be exchangeable: that is the opinion on the unexplained portion of the undercoverage rate for Ontario would be the same as for Prince Edward Island. The final stage specifies the prior distributions of the national undercoverage rate and the unknown variance.

More formally, the hierarchical model is given by:

- (i) The sampling model: $[y_i|\theta_i] \sim N(\theta_i, \psi_i^2)$, $i = 1, \dots, n$, where ψ_i^2 's are known sampling variance;
- (ii) The population model: $[\theta_i|r, \sigma^2] \sim t(r + c_i, \sigma^2, \eta_i)$, where $c_i = \alpha_i - \sum \alpha_k p_k$, σ^2 is the unknown variance component. and η_i 's are the known degrees of freedom associated with the t distributions;
- (iii) Prior distributions: $r \sim N(r_0, \psi_0)$ where r_0 and ψ_0 are known, and $\sigma^2 \sim \Gamma(\frac{1}{2}a, \frac{1}{2}b)$ where a and b are known ($a > 0, b > 0$).

The following comments can be made about the assumed Bayesian model. First, the assumed sampling model is typical of the standard models used in sampling surveys: the assumption of normality, given the large sample size used in coverage studies allows the Central Limit Theorem to be invoked. Secondly, the prior distributions are selected to be proper distributions with known parameters. So we do

not face any problems of improper posteriors (Hobert and Cassella, 1996).

Finally, the usual assumption concerning the true undercoverage rates as being normally distributed is quite restrictive. In particular, the assumption that the variability is the same for all provinces is unlikely to be true. Datta and Lahiri (1995) suggested the t distribution with different degrees of freedom for each area essentially attempting to add an amount of "robustification" to the usual normality assumption. In addition, they recommended that the degrees of freedom, η_i , be set to one for the largest and smallest observations. This corresponds to a Cauchy distribution which falls between a uniform and a normal priors. They also suggested a moderate value of the degrees of freedom for the other observations.

The population model (ii) can be represented in a form more useful for implementing the Gibbs sampler. By using the following lemma, the population model

$$\theta_i | r, \sigma^2 \sim t(r + c_i, \sigma^2, \eta_i)$$

can be represented as

$$\theta_i | r, \sigma^2, u_i \sim N(r + c_i, \frac{1}{u_i}),$$

where u_i is a new variable distributed as

$$u_i | \sigma^2 \sim \Gamma(\frac{1}{2}\eta_i, \frac{1}{2}\sigma^2\eta_i).$$

Lemma: If $x | \lambda \sim N(\mu, \lambda^{-1})$, $\lambda \sim \Gamma(\frac{1}{2}\eta, \frac{1}{2}\sigma^2\eta)$, then marginally $x \sim t(\mu, \sigma^2, \eta)$.

Proof: Straightforward.

Finally, the complete hierarchical Bayes model can now be obtained as:

$$y_i | \theta_i \sim N(\theta_i, \psi_i^2) \tag{3.6}$$

$$\theta_i | r, \sigma^2, u_i \sim N(r + c_i, \frac{1}{u_i}) \quad (3.7)$$

$$u_i | \sigma^2 \sim \Gamma(\frac{1}{2}\eta_i, \frac{1}{2}\sigma^2\eta_i) \quad (3.8)$$

$$r \sim N(r_0, \psi_0) \quad (3.9)$$

$$\sigma^2 \sim \Gamma(\frac{1}{2}a, \frac{1}{2}b) \quad (3.10)$$

where $c_i = \alpha_i - \sum \alpha_k p_k$, and values of α_i , p_i and η_i are prespecified. The problem now is to obtain the posterior mean and posterior variance of the true undercoverage rate for each province.

3.2.4 Gibbs sampling inference

For model (3.6) - (3.10), let $\theta = (\theta_1, \dots, \theta_n)^T$, $Y = (y_1, \dots, y_n)^T$ and $u = (u_1, \dots, u_n)^T$. Then we have variable blocks $(Y, \theta, r, u, \sigma^2)$. To implement the Gibbs sampler, the following full conditional distributions are needed:

(1) For $i = 1, \dots, n$,

$$[\theta_i | Y, r, u, \sigma^2] \sim N\left(\frac{\psi_i^{-2}y_i + (r + c_i)u_i}{\psi_i^{-2} + u_i}, (\psi_i^{-2} + u_i)^{-1}\right) \quad (3.11)$$

(2)

$$[r | Y, \theta, u, \sigma^2] \sim N(\phi^{-1}(\psi_0^{-2}r_0 + \sum u_k(r_k - c_k)), \phi^{-1}) \quad (3.12)$$

where $\phi = \psi_0^{-2} + \sum_{k=1}^n u_k$

(3) For $i = 1, \dots, n$,

$$[u_i | Y, \theta, r, \sigma^2] \sim \Gamma\left(\frac{1}{2}(\eta_i + 1), \frac{1}{2}(\sigma^2\eta_i + (\theta_i - r - c_i)^2)\right) \quad (3.13)$$

(4)

$$[\sigma^2|Y, \theta, r, u] \sim \Gamma(\frac{1}{2}(a + \sum \eta_k), \frac{1}{2}(b + \sum \eta_k u_k)) \quad (3.14)$$

We are mainly interested in estimating θ_i . From (3.11), the Rao-Blackwellized estimator of θ_i is obtained as

$$\hat{\theta}_i = \frac{1}{G} \sum_{k=1}^G \frac{\psi_i^{-2} y_i + (r^{(k)} + c_i) u_i^{(k)}}{\psi_i^{-2} + u_i^{(k)}}. \quad (3.15)$$

and the posterior variance of θ_i is estimated by

$$\begin{aligned} \hat{V}(\theta_i) &= \frac{1}{G} \sum_{k=1}^G (\psi_i^{-2} + u_i^{(k)})^{-1} + \frac{1}{G} \sum_{k=1}^G \left(\frac{\psi_i^{-2} y_i + (r^{(k)} + c_i) u_i^{(k)}}{\psi_i^{-2} + u_i^{(k)}} \right)^2 \\ &\quad - \left(\frac{1}{G} \sum_{k=1}^G \frac{\psi_i^{-2} y_i + (r^{(k)} + c_i) u_i^{(k)}}{\psi_i^{-2} + u_i^{(k)}} \right)^2, \end{aligned} \quad (3.16)$$

where $\{r^{(k)}, u_i^{(k)}, k = 1, \dots, G\}$ are samples generated from the Gibbs sampler.

We implement the Gibbs sampler using BUGS program (Spiegelhalter, et al., 1995). The Gibbs sampler was run for 6,000 iterations. A “burn-in” period of the first 1,000 iterations was discarded. The last 5,000 simulations were kept for estimation and model analysis. Convergence diagnostics was checked by using the CODA Splus program (Best, et al., 1995).

3.3 Census Undercoverage Estimation

3.3.1 Description of the survey estimates

Since 1966, the Reverse Record Check (RRC) has been used by Statistics Canada to measure the gross number of persons missed by the Census. Starting 1991, an Overage Study was conducted to measure the gross number of persons erroneously included in the Census. In 1991, for the first time, the RRC results were combined

with those of the Overcoverage Study to calculate the net undercoverage. Through the analysis of the results of these coverage studies, the collection methodology is adjusted in order to improve coverage in the succeeding Census.

The basic approach of RRC used by the coverage studies is to use several different sampling frames, which are independent lists of persons who should have been included in the Census. These lists are compiled from persons included in the previous Census, persons missed by the Census and persons new to Canada since the last Census. These new persons include the newborns, landed immigrants and non-permanent residents. There are some minor exclusions to these lists, for instance Canadians living outside the country during the previous Census are not included. Sampling was carried out independently within each frame. The sample design varied from frame to frame, depending on the nature of the list used. The sampling rates within frames were not uniform. Higher rates were used in certain subgroups for which high undercoverage was expected, in order to improve the sample design. In 1991, a total of 55,912 persons were selected to estimate the gross undercoverage in the Census. The strategy of Overcoverage Study is basically similar to that used for the Reverse Record Check. More details on the coverage studies can be found in Germain and Julien (1993).

Net undercoverage is the difference between gross undercoverage and overcoverage. The production of estimates of overcoverage from the 1991 Overcoverage Study, combined with estimates of undercoverage from the 1991 Reverse Record Check, has made it possible to produce estimates of net undercoverage for the first time. The direct survey estimates y_i used in the hierarchical Bayes model are the results of the

coverage studies for the provincial net undercoverage in the Census. These direct estimates can be made more accurate by applying model based methods such as the hierarchical Bayes method.

3.3.2 Estimation and results

For the hierarchical Bayes estimation of the Census undercoverage rate, model (3.6)-(3.10) is used. The values of α_i for the fixed effects part and the degrees of freedom η_i for the t distribution need to be specified. Four different set of values were taken when specifying the α_i 's. Table 3.1 presents the specification of the values of α corresponding to each model.

Model 1 assumes that all provinces have exchangeable errors when sampled from the national undercoverage rate. This model implies that all $\alpha_i = 0$. Model 2 assumes that Ontario has an undercoverage rate that is 1% higher and PEI has an undercoverage rate that is 1.5% lower than the other eight provinces. This model is accounting for the fact that it is becoming relatively difficult to conduct a Census in Ontario, and in particular in Toronto area because of its size, diversity and complexity. On the other hand, this model allows for the apparent ease in which a Census can be conducted in PEI due to small size and a stable and relatively homogeneous population. Hence we have $\alpha_2 = -0.015$, $\alpha_6 = 0.01$, and for $i \neq 2$ or 6 , $\alpha_i = 0$. For Model 3, values of α_i 's are determined by using the gross undercoverage results for 1986. That is, we use the 1986 result as our prior information on the fixed effects. For Model 4, values of α_i 's are determined by 1991 Population Projection Estimates.

The degrees of freedom η_i of t distributions in the population model also need to

be specified. Three different sets of values were assigned to η_i 's within each model. First, in order to provide a benchmark series, we have assumed η_i to be very large, say $\eta_i = 200$ for all provinces. This in fact assumes that the population model has a normal distribution for each province. Then following the suggestion of Datta and Lahiri (1994), we have assigned one degree of freedom (a Cauchy distribution) to the minimum and maximum observations, i.e, PEI and Ontario. For the other provinces, we have assumed 5 or 15 degrees of freedom.

We used proper priors for the unknown hyperparameters r and σ^2 . For the national undercoverage rate r , we assumed that $r \sim N(0.02865, 1)$ which is a normal distribution with mean equal to the observed national undercoverage rate in 1991 and relatively large variance. We assumed $\sigma^2 \sim \Gamma(0.001, 0.001)$ to reflect our vague prior knowledge about this variance component. The results of the posterior estimates for the various models are displayed in Table 3.2. For each model, first we considered the normal distribution for the population model for all provinces. Secondly, we considered Cauchy distribution for PEI and Ontario and for the rest of the eight provinces, we used t distribution with 5 and 15 degrees of freedom.

Model 1 shows that all provinces have moved toward the national undercoverage rate of 2.86 %. The provinces close to the national rate, such as BC and Quebec move very little while the extreme provinces such as PEI, New Brunswick and Ontario move the most, especially when the population model is normal. The results of t distribution with 5 or 15 degrees of freedom are similar. Under Model 2, the estimates are similar for different η_i , which indicates that Model 2 produces more stable estimates. This seems to be due to the fixed effects component of the popula-

tion model. Model 3 shows a big increase in BC due to the fixed effect $\alpha_{10} = 0.0127$ which is contrary to what was observed in 1991. The result of Model 4 is very similar to that of Model 1 since all the values of α_i 's for Model 4 are relatively small and close to 0.

Table 3.3 shows the provincial efficiency gains of the hierarchical Bayes estimates over the direct survey estimates y_i . The efficiency is defined as the ratio of the sampling variance to the posterior variance. As to the efficiency comparison, Model 2 has the largest efficiency gain among the four models. The average efficiency of Model 2 is about 1.21. Model 1 has the smallest efficiency with an average about 1.09. Model 3 and Model 4 have the same average efficiency 1.13. For Model 1 and Model 4, the most notable result is the slight loss in efficiency experienced in PEI when the population model is assumed to be a t distribution. For Model 2, all provinces show a gain in efficiency; in particular, New Brunswick, Manitoba, Ontario and Nova Scotia have large gains in efficiency. From the efficiency comparison, the posterior estimate of undercoverage rate under Model 2 has the smallest variability among the four models.

3.3.3 Bayesian model determination

As discussed in Chapter 2, the hierarchical Bayes method and Markov Chain Monte Carlo methods enable us to fit various models for data analysis. In practice, responsible data analysis should address the issue of model determination. In this section, we will perform residual analysis using the cross-validation predictive densities for evaluation of model adequacy for the Census undercoverage models. Then we will use the conditional predictive ordinate(CPO) based on cross-validation predictive

densities for Bayesian model choice. Detailed definitions and formulas can be found in Chapter 2. These Bayesian methods are in contrast to traditional statistical methods, where two models cannot be easily compared unless one is a special case of the other, i.e., the models are nested. See Gilks et al. (1996) for more discussions and examples.

Residual analysis

For checking adequacy of the Census undercoverage models, we calculated the cross-validation residual for each y_i as

$$d_i = y_{i,obs} - E(y_i | Y_{(i),obs}), \quad (3.17)$$

where $Y = (y_1, \dots, y_n)$, and $Y_{(i)}$ denotes all elements of Y except y_i . To compute $E(y_i | Y_{(i),obs})$, we write

$$E(y_i | Y_{(i),obs}) = E_1 E_2(y_i | \theta, Y_{(i),obs}) = E_1(h_i(\theta, Y_{(i),obs})) \quad (3.18)$$

where E_1 denotes a conditional expectation over θ given $Y_{(i),obs}$, E_2 denotes a conditional expectation over y_i given θ and $Y_{(i),obs}$, and $h_i(\theta, Y_{(i),obs}) = E_2(y_i)$. If y_i 's are conditionally independent, then h_i is free of $Y_{(i),obs}$ and only depends on θ . Then we have

$$\begin{aligned} E_1(h_i(\theta)) &= \int h_i(\theta) f(\theta | Y_{(i),obs}) d\theta \\ &= f(y_{i,obs} | Y_{(i),obs}) \int \frac{h_i(\theta)}{f(y_{i,obs} | \theta)} f(\theta | Y_{obs}) d\theta \end{aligned} \quad (3.19)$$

Thus a straightforward Monte Carlo integration for (3.19) yields

$$\hat{E}(y_i | Y_{(i),obs}) = \hat{f}(y_{i,obs} | Y_{(i),obs}) \cdot \frac{1}{m} \sum_{j=1}^m \frac{h_i(\theta_j^*)}{f(y_{i,obs} | \theta_j^*)}. \quad (3.20)$$

where $\hat{f}(y_{i,obs}|Y_{(i),obs})$ is actually the estimated CPO value for the i th observation y_i , which is given in (3.23). Since $E(y_i|\theta_i) = \theta_i$, then $h_i(\theta) = \theta_i$, and $E(y_i|Y_{(i),obs})$ is estimated as

$$\hat{E}(y_i|Y_{(i),obs}) = \hat{f}(y_{i,obs}|Y_{(i),obs}) \cdot \frac{1}{m} \sum_{j=1}^m \frac{\theta_{ij}^*}{f(y_{i,obs}|\theta_{ij}^*)}, \quad (3.21)$$

where $\{\theta_{ij}^*, j = 1, \dots, m\}$ is the simulated samples of the posterior distribution of θ_i from the Gibbs sampler. We use sum of absolute residuals, denoted as SAR, as the total residual measurement for the model adequacy checking. Thus SAR is defined as

$$SAR = \sum_{i=1}^n |y_{i,obs} - \hat{E}(y_i|Y_{(i),obs})| \quad (3.22)$$

Table 3.4 shows the cross-validation predictive residuals and their respective SAR under each model. Model 2 has the smallest sum of absolute residuals (SAR), and Model 1 has the largest SAR. Model 2 is the best fitting model according to the residual comparison. Within each model, population model with t distribution has smaller SAR than the one with normal distribution ($\eta_i = 200$), which indicates that t distribution is better than the normal distribution population model for model fitting.

Bayesian model choice

For the Census undercoverage models, y_i are conditionally independent. Thus by the result in Chapter 2, the estimate of CPO_i for each province is given by

$$\widehat{CPO}_i = \hat{f}(y_{i,obs}|Y_{(i),obs}) = \frac{1}{\frac{1}{m} \sum_{j=1}^m \frac{1}{f(y_{i,obs}|\theta_{ij}^*)}}. \quad (3.23)$$

The estimated $\widehat{CPO}_i$ was calculated based on $m = 5,000$ simulated samples $\{\theta_{ij}^*, j = 1, \dots, m\}$ from the posterior of θ_i for each province under different models. A suitable summary measure for CPO might be the log of the product of the CPOs for each datapoint. Thus we define

$$\log L = \log(\prod_{i=1}^n \widehat{CPO}_i). \quad (3.24)$$

Note that $\log L$ is just the log likelihood of all the observed data. Large value of $\log L$ suggests a more likely model. Table 3.5 shows the CPO values and $\log L$ under each model.

Model 2 generally has the largest CPO values for all provinces except New Brunswick (4th observation). Also Model 2 has the largest value of log likelihood among the four models. The log likelihood of Model 2 is around 39.24. Whereas for the other three models the value of $\log L$ is less than 38, Model 1 has the smallest $\log L$. Values of CPO and $\log L$ suggest that Model 2 is the most likely model for the data.

For pointwise comparison of the four models, we calculated the average CPO values within each model. Table 3.6 shows the average CPO values for each model. We also present a CPO plot for the four models in Figure 3.1. Since larger CPO values are indicative of better model fit, Model 2 is the best one among the four models. Model 2 is better than Model 1 at all the observations except the 4th one. Model 2 is better than Model 3 at all the observations except the 3rd and 8th ones. And Model 2 is better than Model 4 at all the observations except the 3rd and 4th ones. The CPO values of the four models are all small for New Brunswick, which indicates that New Brunswick could be an outlier. Since New Brunswick is a small

province without large cities, its estimated undercoverage rate should not be as high as 3.25%. Overall most of the data support Model 2. Model 1 is the worst model. Model 3 performs very badly at the 10th observation (BC). And Model 4 performs badly at the 8th observation (SASK).

3.4 Conclusion

Hierarchical Bayes model-based estimates have been used to estimate Canadian Census provincial undercoverage rates. Four different models are considered by specifying different prior knowledge on the fixed effects in the population model within the hierarchical Bayes structure. Hierarchical Bayes estimates are obtained by Gibbs sampling procedure and have been shown some improvement over the direct survey estimates. Model 2 provides more accurate estimates than the other three models.

Bayesian model checking and choice are also illustrated by using the cross-validation predictive densities. From the residual analysis and CPO comparison, Model 2 is the most likely model and most of the data support Model 2.

Further work need to be done to evaluate the fixed effects in the population model within the hierarchical Bayes model since the fixed effects play an important role in the procedure of estimation. In our model, sampling variances are assumed to be known and fixed. A more complicated hierarchical model would include a model for the sampling variances. Finally, to overcome the limitations of the current models mentioned in Section 3.2.1, we may consider non-linear mixed models; see Chapter 9 for more details.

Specification of α for Four Models				
Province	Model 1	Model 2	Model 3	Model 4
(1) NFLD	0	0	-0.0129	0.0023
(2) PEI	0	-0.015	-0.0107	0
(3) NS	0	0	-0.0106	-0.0063
(4) NB	0	0	-0.0050	0.0005
(5) QUE	0	0	-0.0030	-0.0038
(6) ONT	0	0.01	0.0022	0.0050
(7) MAN	0	0	-0.0027	-0.0031
(8) SASK	0	0	-0.0083	0.0001
(9) ALTA	0	0	-0.0021	-0.0048
(10) BC	0	0	0.0127	-0.0015

Table 3.1: Model specification

Province	y_i	Model 1			Model 2		
		$\eta_i = 200$	$\eta_i = 5$	$\eta_i = 15$	$\eta_i = 200$	$\eta_i = 5$	$\eta_i = 15$
NFLD	0.0199	0.0206	0.0208	0.0208	0.0208	0.0208	0.0208
PEI	0.0093	0.0107	0.0100	0.0100	0.0093	0.0094	0.0094
NS	0.0189	0.0199	0.0202	0.0202	0.0203	0.0202	0.0202
NB	0.0325	0.0317	0.0315	0.0315	0.0301	0.0302	0.0302
QUE	0.0261	0.0261	0.0261	0.0261	0.0259	0.0259	0.0259
ONT	0.0364	0.0357	0.0355	0.0355	0.0360	0.0360	0.0360
MAN	0.0186	0.0198	0.0200	0.0201	0.0201	0.0200	0.0201
SASK	0.0180	0.0190	0.0192	0.0193	0.0193	0.0193	0.0193
ALTA	0.0200	0.0205	0.0207	0.0207	0.0207	0.0207	0.0207
BC	0.0273	0.0273	0.0274	0.0274	0.0269	0.0269	0.0269

Province	y_i	Model 3			Model 4		
		$\eta_i = 200$	$\eta_i = 5$	$\eta_i = 15$	$\eta_i = 200$	$\eta_i = 5$	$\eta_i = 15$
NFLD	0.0199	0.0193	0.0192	0.0193	0.0210	0.0212	0.0213
PEI	0.0093	0.0102	0.0102	0.0101	0.0109	0.0101	0.0100
NS	0.0189	0.0186	0.0186	0.0186	0.0192	0.0194	0.0194
NB	0.0325	0.0304	0.0304	0.0304	0.0317	0.0315	0.0315
QUE	0.0261	0.0260	0.0260	0.0260	0.0259	0.0259	0.0259
ONT	0.0364	0.0357	0.0356	0.0355	0.0361	0.0359	0.0358
MAN	0.0186	0.0198	0.0199	0.0198	0.0195	0.0198	0.0198
SASK	0.0180	0.0182	0.0183	0.0183	0.0192	0.0195	0.0195
ALTA	0.0200	0.0206	0.0207	0.0206	0.0202	0.0204	0.0203
BC	0.0273	0.0286	0.0284	0.0286	0.0273	0.0273	0.0273

Table 3.2: Direct and Posterior Provincial Undercoverage Estimates

Province	y_i	Model 1			Model 2		
		$\eta_i = 200$	$\eta_i = 5$	$\eta_i = 15$	$\eta_i = 200$	$\eta_i = 5$	$\eta_i = 15$
NFLD	1	1.08	1.08	1.10	1.19	1.18	1.19
PEI	1	1.04	0.98	0.96	1.21	1.24	1.24
NS	1	1.10	1.10	1.14	1.26	1.25	1.25
NB	1	1.18	1.20	1.20	1.34	1.31	1.30
QUE	1	1.06	1.06	1.06	1.11	1.11	1.11
ONT	1	1.07	1.03	1.03	1.20	1.25	1.25
MAN	1	1.13	1.11	1.14	1.29	1.25	1.26
SASK	1	1.12	1.07	1.10	1.25	1.20	1.21
ALTA	1	1.08	1.07	1.09	1.17	1.16	1.16
BC	1	1.07	1.09	1.11	1.15	1.14	1.15
Average	1	1.09	1.08	1.09	1.22	1.21	1.21

Province	y_i	Model 3			Model 4		
		$\eta_i = 200$	$\eta_i = 5$	$\eta_i = 15$	$\eta_i = 200$	$\eta_i = 5$	$\eta_i = 15$
NFLD	1	1.13	1.15	1.15	1.09	1.09	1.10
PEI	1	1.12	1.05	1.04	1.04	0.99	0.98
NS	1	1.18	1.25	1.22	1.15	1.25	1.22
NB	1	1.22	1.13	1.15	1.22	1.25	1.26
QUE	1	1.09	1.08	1.08	1.08	1.08	1.09
ONT	1	1.11	1.09	1.08	1.10	1.16	1.14
MAN	1	1.19	1.16	1.18	1.17	1.17	1.18
SASK	1	1.19	1.19	1.19	1.13	1.12	1.12
ALTA	1	1.12	1.11	1.08	1.11	1.10	1.11
BC	1	1.05	1.05	1.06	1.08	1.09	1.09
Average	1	1.14	1.13	1.12	1.12	1.13	1.13

Table 3.3: Provincial Efficiency Comparison

Province	Model 1			Model 2		
	$\eta_i = 200$	$\eta_i = 5$	$\eta_i = 15$	$\eta_i = 200$	$\eta_i = 5$	$\eta_i = 15$
NFLD	-0.00261	-0.00269	-0.00236	-0.00223	-0.00235	-0.00213
PEI	-0.00622	-0.00223	-0.00296	-0.00056	0.00125	0.00121
NS	-0.00602	-0.00654	-0.00496	-0.00478	-0.00492	-0.00408
NB	0.00304	0.00297	0.00313	0.00710	0.00684	0.00671
QUE	0.00039	-0.00063	-0.00139	0.00086	0.00053	-0.00013
ONT	0.00556	0.00403	0.00504	0.00267	0.00031	0.00187
MAN	-0.00520	-0.00509	-0.00503	-0.00460	-0.00430	-0.00444
SASK	-0.00422	-0.00622	-0.00423	-0.00436	-0.00413	-0.00392
ALTA	-0.00182	-0.00315	-0.00374	-0.00202	-0.00266	-0.00294
BC	-0.00157	-0.00104	-0.00029	0.00063	0.00100	0.00123
SAR	0.03665	0.03459	0.03313	0.02981	0.02829	0.02866

Province	Model 3			Model 4		
	$\eta_i = 200$	$\eta_i = 5$	$\eta_i = 15$	$\eta_i = 200$	$\eta_i = 5$	$\eta_i = 15$
NFLD	0.00440	0.00321	0.00329	-0.00439	-0.00447	-0.00474
PEI	-0.00403	-0.00174	-0.00284	-0.00714	-0.00436	-0.00457
NS	0.00025	0.00034	0.00126	-0.00241	-0.00248	-0.00232
NB	0.00741	0.00743	0.00762	0.00296	0.00267	0.00268
QUE	0.00115	0.00020	-0.00063	0.00146	0.00052	-0.00021
ONT	0.00504	0.00367	0.00374	0.00341	0.00237	0.00259
MAN	-0.00456	-0.00412	-0.00395	-0.00367	-0.00378	-0.00368
SASK	-0.00065	-0.00269	-0.00068	-0.00488	-0.00692	-0.00705
ALTA	-0.00196	-0.00281	-0.00329	-0.00049	-0.00157	-0.00179
BC	-0.00674	-0.00613	-0.00550	-0.00099	-0.00094	-0.00102
SAR	0.03619	0.03234	0.03280	0.03180	0.03008	0.03065

Table 3.4: Bayesian Cross-Validation Predictive Residuals

Province	Model 1			Model 2		
	$\eta_i = 200$	$\eta_i = 5$	$\eta_i = 15$	$\eta_i = 200$	$\eta_i = 5$	$\eta_i = 15$
NFLD	42.367	44.923	46.706	50.750	52.647	54.423
PEI	30.089	32.044	32.803	64.949	57.336	57.796
NS	32.098	30.433	34.184	41.496	40.405	42.454
NB	36.685	35.664	36.766	31.024	30.577	30.343
QUE	63.113	68.187	66.110	70.908	78.457	73.953
ONT	34.103	34.191	33.293	52.645	56.609	55.174
MAN	34.005	33.720	32.479	41.706	41.947	40.644
SASK	39.045	38.213	40.375	44.679	42.690	45.528
ALTA	54.489	45.423	42.974	60.618	55.282	54.416
BC	52.515	56.293	60.224	61.159	63.529	68.126
$\log L$	37.046	37.024	37.205	39.234	39.185	39.270
Province	Model 3			Model 4		
	$\eta_i = 200$	$\eta_i = 5$	$\eta_i = 15$	$\eta_i = 200$	$\eta_i = 5$	$\eta_i = 15$
NFLD	40.703	46.127	45.030	37.644	38.014	39.108
PEI	40.733	39.569	37.463	28.842	30.373	32.690
NS	47.723	50.932	48.218	43.204	45.427	46.994
NB	28.370	25.021	25.835	38.647	38.664	40.166
QUE	64.182	75.546	68.845	62.049	70.813	69.850
ONT	38.922	42.177	40.853	43.292	46.868	48.763
MAN	38.651	39.018	37.254	39.097	40.692	38.683
SASK	50.125	45.336	49.177	29.717	27.120	30.928
ALTA	52.888	48.532	46.596	55.282	54.853	53.562
BC	26.087	28.553	33.975	54.614	59.264	60.178
$\log L$	37.246	37.438	37.380	37.377	37.723	37.995

Table 3.5: CPO and log Likelihood Values

Pointwise CPO Estimates				
Province	Model 1	Model 2	Model 3	Model 4
NFLD	44.665	52.607	43.953	38.255
PEI	31.645	60.027	39.255	30.635
NS	32.238	41.451	48.957	45.208
NB	36.371	30.648	26.408	39.159
QUE	65.803	74.439	69.524	67.571
ONT	33.862	54.809	40.650	46.308
MAN	33.401	41.432	38.307	39.491
SASK	39.211	44.299	48.212	29.255
ALTA	47.628	56.772	49.338	54.566
BC	56.344	64.271	29.538	58.019

Table 3.6: Pointwise CPO Comparison

Figure 3.1: CPO Plot of Census Undercoverage Models

Chapter 4

Hierarchical Bayes Small Area Estimation Using Multi-level Models

4.1 Introduction

Battese et al. (1981,1988) proposed and applied a nested error regression model to provide model-based small area estimates. The model takes the form:

$$Y_i = X_i\beta + v_{0i}1_{n_i} + e_i, \quad (4.1)$$

where Y_i is the $n_i \times 1$ vector of observations $y_{ij}(j = 1, \dots, n_i)$ associated with the sampled units in the i th small area, $i = 1, \dots, m$, X_i is the $n_i \times p$ matrix of explanatory variables, $1_{n_i} = (1, \dots, 1)^T$ is the unit vector of length n_i , β is a set of p fixed regression parameters, v_{0i} are independent area effects with $E(v_{0i}) = 0$ and $V(v_{0i}) = \sigma_v^2$. The e_i 's are assumed to be independent random error vectors with $E(e_i) = 0$ and $V(e_i) = \sigma_e^2 I_{n_i}$, where I_{n_i} is the $n_i \times n_i$ identity matrix. v_{0i} and e_i are also assumed to be independent. For the whole population, model (4.1) applies with n_i replaced by N_i , the small area population size.

Holt and Moura (1993, 1999) extended the above framework to a multi-level model by introducing random regression coefficients and then relating them to area-level explanatory variables to explain some of the between small area variation. The model can be stated as follows:

$$\begin{aligned} Y_i &= X_i\beta_i + e_i \\ \beta_i &= Z_i\gamma + v_i \end{aligned} \tag{4.2}$$

where Z_i is the $p \times q$ design matrix of small area level variables, γ is a $q \times 1$ vector of fixed coefficients, and $v_i = (v_{i1}, \dots, v_{ip})^T$ is a $p \times 1$ vector of random effects for the i th small area. The v_i 's are independent across small areas and have a joint distribution within each small area with $E(v_i) = 0$ and $V(v_i) = \Phi$, where the variance covariance matrix Φ is unknown and has a general form

$$\Phi = \begin{bmatrix} \phi_1^2 & \phi_{12} & \cdots & \phi_{1p} \\ \phi_{21} & \phi_2^2 & \cdots & \phi_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{p1} & \phi_{p2} & \cdots & \phi_p^2 \end{bmatrix}.$$

Note that model (4.2) effectively integrates the use of unit-level and area-level covariates into a single model. Holt and Moura (1993, 1999) extended Prasad and Rao's (1990) framework to the above multi-level model to get the best linear unbiased predictor (BLUP) of the small area mean $\mu_i = \bar{X}_i^T \beta_i$ assuming that N_i is large, where $\bar{X}_i$ is the $p \times 1$ vector of population means of the auxiliary variables for the i th small area. Holt and Moura (1993) also obtained the empirical BLUP (EBLUP) and a second order approximation to the mean squared error (MSE) of EBLUP for the multi-level model.

In this Chapter, we study the multi-level model (4.2) in a hierarchical Bayes framework and extend the model to more general multi-level models which allow fixed unequal error variances or random error variances. The small area mean μ_i is estimated by its posterior mean and its precision is measured by its posterior variance. Posterior variance automatically takes into account the extra uncertainty associated with the hyperparameters in the multi-level model. We use the Gibbs sampling method to compute the hierarchical Bayes estimates and the associated posterior variances. In Sections 4.2, we present the hierarchical Bayes multi-level models with different assumptions on error variances and the corresponding Gibbs sampling inferences. In Section 4.3, we illustrate our methodology and also study model selection, model comparison and sensitivity analysis by employing data on household incomes in some counties (small areas) of Brazil.

4.2 Hierarchical Bayes Multi-level Models

4.2.1 Equal error variance model

A hierarchical Bayes representation of the multi-level model (4.2) is given by:

Model 1:

(i) Conditional on β_i and σ_e^2 , y_{ij} 's are independent with

$$y_{ij}|\beta_i, \sigma_e^2 \sim N(x_{ij}^T\beta_i, \sigma_e^2), \quad (i = 1, \dots, m; j = 1, \dots, n_i); \quad (4.3)$$

(ii) Conditional on γ and Φ , β_i 's are independent with

$$\beta_i|\gamma, \Phi \sim N_p(Z_i\gamma, \Phi), \quad (i = 1, \dots, m); \quad (4.4)$$

To complete our Bayesian model specification, we adopt the prior distributions as follows:

- (iii) Marginal prior distributions: $\gamma \sim N_q(0, D)$, $\tau_e \sim G(a, b)$ and $\Omega \sim W_p(\alpha, R)$, where $\tau_e = \sigma_e^{-2}$, $\Omega = \Phi^{-1}$, and D , a , b , α and R are known.

In step (iii) of Model 1, $G(a, b)$ denotes a gamma distribution with density given by $f(x) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-xb}$, $a > 0$, $b > 0$, $x \geq 0$, and $W_p(\alpha, R)$ is a Wishart distribution with density function

$$f(X) = \frac{|R|^{\frac{a}{2}}}{2^{\frac{ap}{2}} \Gamma_p(\frac{a}{2})} |X|^{\frac{a-p-1}{2}} \exp\{-\frac{1}{2} \text{tr}(RX)\}.$$

where $X > 0$, $R > 0$ and $\Gamma_p(\alpha)$ is multivariate gamma function defined as $\Gamma_p(\alpha) = \pi^{\frac{p(p-1)}{4}} \prod_{j=1}^p \Gamma(\alpha + \frac{1}{2}(1-j))$.

Remark 1.1: The prior distributions in step (iii) are conjugate with the sampling and population distributions given by (4.3) and (4.4) in the sense that they lead to full conditional distributions for γ , τ_e and Ω that are again normal, gamma and Wishart distribution, respectively.

Remark 1.2: It is important to note that we have used proper priors on all the unknown parameters to ensure that all the posterior distributions are proper (Hobert and Casella, 1996). Hence we do not face the problem of some posteriors being improper. Values for the parameters of the priors (i.e., hyperparameters) are chosen to reflect a fairly vague knowledge of the prior distributions. Details will be given in Section 4.3 on data analysis.

Remark 1.3: In Model 1, we assume equal error variance σ_e^2 for all small areas. In practice, however, sampling error variances could be different for different small areas. A more general model should allow possibly different error variances.

we are interested in finding the posterior distributions of β_i 's given the data $Y = (\{y_{ij}\}, i = 1, \dots, m; j = 1, \dots, n_i)$, and in particular in finding the posterior estimates of small area means $\mu_i = \bar{X}_i^T \beta_i$, which depend on the estimates of β_i . Direct evaluation of the joint posterior distribution involves high-dimensional numerical integration, and is not computationally feasible. Thus we use the Gibbs sampling method (Gelfand and Smith, 1990) to generate samples from the joint posterior distributions. The full conditional distributions for Gibbs sampling under Model 1 are given by:

(i) For $i = 1, \dots, m$, $[\beta_i | Y, \gamma, \Omega, \tau_e] \stackrel{ind}{\sim}$

$$N_p((\tau_e X_i^T X_i + \Omega)^{-1}(\tau_e X_i^T Y_i + \Omega Z_i \gamma), (\tau_e X_i^T X_i + \Omega)^{-1})$$

(ii) $[\gamma | Y, \beta, \Omega, \tau_e] \sim$

$$N_q((\sum_{i=1}^m Z_i^T \Omega Z_i + D^{-1})^{-1}(\sum_{i=1}^m Z_i^T \Omega \beta_i), (\sum_{i=1}^m Z_i^T \Omega Z_i + D^{-1})^{-1})$$

(iii) $[\Omega | Y, \beta, \gamma, \tau_e] \sim W_p(\alpha + m, R + \sum_{i=1}^m (\beta_i - Z_i \gamma)(\beta_i - Z_i \gamma)^T)$

(iv) $[\tau_e | Y, \beta, \gamma, \Omega] \sim G(a + \frac{1}{2} \sum_{i=1}^m n_i, b + \frac{1}{2} (\sum_{i=1}^m (Y_i - X_i \beta_i)^T (Y_i - X_i \beta_i))$

Suppose a sample of size G is obtained as $\{\beta_i^{(k)}, i = 1, \dots, m; \gamma^{(k)}; \Omega^{(k)}; \tau_e^{(k)}\}, k = 1, \dots, G$. To obtain an estimator of the posterior mean of β_i , one can use the sample mean of the $\{\beta_i^{(k)}\}$. Since β_i has a closed form full conditional distribution, we can use the sample mean of the conditional expectations $\{E[\beta_i | Y, \gamma^{(k)}, \Omega^{(k)}, \tau_e^{(k)}]\}$ to improve our estimation, since $E(\beta_i | Y) = E(E(\beta_i | Y, \gamma, \Omega, \tau_e))$ and $Var(\beta_i | Y) \geq Var(E(\beta_i | Y, \gamma, \Omega, \tau_e))$. This modification is based on the Rao-Blackwell theorem and the corresponding estimator is the so-called Rao-Blackwellized estimator (Gelfand

and Smith 1990, 1991). Thus we have the following two estimators for β_i :

$$\hat{\beta}_i^{(E)} = \frac{1}{G} \sum_{k=1}^G \beta_i^{(k)} \quad (4.5)$$

and

$$\begin{aligned} \hat{\beta}_i^{(RB)} &= \frac{1}{G} \sum_{k=1}^G E(\beta_i | Y, \gamma^{(k)}, \Omega^{(k)}, \tau_e^{(k)}) \\ &= \frac{1}{G} \sum_{k=1}^G (\tau_e^{(k)} X_i^T X_i + \Omega^{(k)})^{-1} (\tau_e^{(k)} X_i^T Y_i + \Omega^{(k)} Z_i \gamma^{(k)}) \end{aligned} \quad (4.6)$$

where $\hat{\beta}_i^{(E)}$ is the empirical estimator and $\hat{\beta}_i^{(RB)}$ is the Rao-Blackwellized estimator. $\hat{\beta}_i^{(RB)}$ is better than $\hat{\beta}_i^{(E)}$ in terms of mean squared error (Gelfand and Smith, 1991).

The corresponding estimators for the small area mean μ_i are given as

$$\hat{\mu}_i^{(E)} = \bar{X}_i^T \hat{\beta}_i^{(E)} = \frac{1}{G} \sum_{k=1}^G \bar{X}_i^T \beta_i^{(k)} \quad (4.7)$$

and

$$\begin{aligned} \hat{\mu}_i^{(RB)} &= \bar{X}_i^T \hat{\beta}_i^{(RB)} \\ &= \frac{1}{G} \sum_{k=1}^G \bar{X}_i^T (\tau_e^{(k)} X_i^T X_i + \Omega^{(k)})^{-1} (\tau_e^{(k)} X_i^T Y_i + \Omega^{(k)} Z_i \gamma^{(k)}) \end{aligned} \quad (4.8)$$

It will be of interest to compute and compare the simulation standard errors of these two estimators; see Section 5.

To obtain the posterior variance of μ_i , we first find the posterior variance of β_i , since $V(\mu_i | Y) = \bar{X}_i^T V(\beta_i | Y) \bar{X}_i$. Note that

$$\begin{aligned} V(\beta_i | Y) &= E(V(\beta_i | Y, \gamma, \Omega, \tau_e)) + V(E(\beta_i | Y, \gamma, \Omega, \tau_e)) \\ &= E(V(\beta_i | Y, \gamma, \Omega, \tau_e)) + E(E(\beta_i | Y, \gamma, \Omega, \tau_e)^2) \\ &\quad - [E(E(\beta_i | Y, \gamma, \Omega, \tau_e))]^2. \end{aligned} \quad (4.9)$$

It follows from (4.9) that the Rao-Blackwellized estimator of the posterior variance of β_i is given by

$$\begin{aligned}
\hat{V}(\beta_i) &= \frac{1}{G} \sum_{k=1}^G V(\beta_i | Y, \gamma^{(k)}, \Omega^{(k)}, \tau_e^{(k)}) + \frac{1}{G} \sum_{k=1}^G [E(\beta_i | Y, \gamma^{(k)}, \Omega^{(k)}, \tau_e^{(k)})]^2 \\
&\quad - [\frac{1}{G} \sum_{k=1}^G E(\beta_i | Y, \gamma^{(k)}, \Omega^{(k)}, \tau_e^{(k)})]^2 \\
&= \frac{1}{G} \sum_{k=1}^G (\tau_e^{(k)} X_i^T X_i + \Omega^{(k)})^{-1} \\
&\quad + \frac{1}{G} \sum_{k=1}^G (\tau_e^{(k)} X_i^T X_i + \Omega^{(k)})^{-1} (\tau_e^{(k)} X_i^T Y_i + \Omega^{(k)} Z_i \gamma^{(k)}) \\
&\quad \times (\tau_e^{(k)} X_i^T Y_i + \Omega^{(k)} Z_i \gamma^{(k)})^T (\tau_e^{(k)} X_i^T X_i + \Omega^{(k)})^{-1} \\
&\quad - \frac{1}{G^2} [\sum_{k=1}^G (\tau_e^{(k)} X_i^T X_i + \Omega^{(k)})^{-1} (\tau_e^{(k)} X_i^T Y_i + \Omega^{(k)} Z_i \gamma^{(k)})] \\
&\quad \times [\sum_{k=1}^G (\tau_e^{(k)} X_i^T X_i + \Omega^{(k)})^{-1} (\tau_e^{(k)} X_i^T Y_i + \Omega^{(k)} Z_i \gamma^{(k)})]^T. \tag{4.10}
\end{aligned}$$

Then the posterior variance of small area mean μ_i is estimated by

$$\hat{V}(\mu_i) = \overline{X}_i^T \hat{V}(\beta_i) \overline{X}_i \tag{4.11}$$

where $\hat{V}(\beta_i)$ is given by (4.10).

The same estimation procedure can be applied to the sampling error variance σ_e^2 . Since $\tau_e = \sigma_e^{-2}$, the empirical estimator of σ_e^2 is

$$\hat{\sigma}_e^{2(E)} = \frac{1}{G} \sum_{k=1}^G \frac{1}{\tau_e^{(k)}} \tag{4.12}$$

and the Rao-Blackwellized estimator of the posterior mean of σ_e^2 is

$$\hat{\sigma}_e^{2(RB)} = \frac{1}{G} \sum_{k=1}^G [b + \frac{1}{2} (\sum_{i=1}^m (Y_i - X_i \beta_i^{(k)})^T (Y_i - X_i \beta_i^{(k)}))] (a + \frac{1}{2} \sum_{i=1}^m n_i - 1)^{-1}. \tag{4.13}$$

The Rao-Blackwellized estimator of the posterior variance of σ_e^2 is given by

$$\begin{aligned}
 \hat{V}(\sigma_e^2) &= \frac{1}{G} \sum_{k=1}^G [b + \frac{1}{2} \sum_{i=1}^m (Y_i - X_i \beta_i^{(k)})^T (Y_i - X_i \beta_i^{(k)})]^2 \\
 &\quad \times (a + \frac{1}{2} \sum_{i=1}^m n_i - 1)^{-2} [(a + \frac{1}{2} \sum_{i=1}^m n_i - 2)^{-1} + 1] \\
 &\quad - \frac{1}{G^2} [\sum_{k=1}^G (b + \frac{1}{2} \sum_{i=1}^m (Y_i - X_i \beta_i^{(k)})^T (Y_i - X_i \beta_i^{(k)})) \\
 &\quad \times (a + \frac{1}{2} \sum_{i=1}^m n_i - 1)^{-1}]^2
 \end{aligned} \tag{4.14}$$

4.2.2 Unequal error variance model

In practice, it is more realistic to allow unequal error variances for the sampling errors. Let σ_i^2 be the true sampling error variance for the i th small area. A straightforward extension of Model 1 leads to the following hierarchical Bayes multi-level unequal error variance model:

Model 2:

(i) Conditional on β_i and σ_i^2 , y_{ij} 's are independent with

$$y_{ij} | \beta_i, \sigma_i^2 \sim N(x_{ij}^T \beta_i, \sigma_i^2), \quad (i = 1, \dots, m; j = 1, \dots, n_i); \tag{4.15}$$

(ii) Conditional on γ and Φ , β_i 's are independent with

$$\beta_i | \gamma, \Phi \sim N_p(Z_i \gamma, \Phi), \quad (i = 1, \dots, m); \tag{4.16}$$

(iii) Marginal prior distributions: $\gamma \sim N_q(0, D)$, $\tau_i \stackrel{ind}{\sim} G(a_i, b_i)$, and $\Omega \sim W_p(\alpha, R)$, where $\tau_i = \sigma_i^{-2}$, $\Omega = \Phi^{-1}$, and D , a_i , b_i , α and R are known.

Remark 2.1: Model 2 reduces to Model 1 when $\sigma_i^2 = \sigma_e^2$ for all i . From a hierarchical Bayes perspective, extension from the equal error variance model to the unequal error variance model is straightforward. Also there is no difficulty in the Gibbs sampling implementation.

Remark 2.2: τ_i 's are assumed to be independent and have prior distributions $G(a_i, b_i)$, where a_i and b_i are known hyperparameters and usually chosen to be very small to reflect a vague knowledge about τ_i 's.

The full conditional distributions for Gibbs sampling under Model 2 are given by:

(i) For $i = 1, \dots, m$, $[\beta_i | Y, \gamma, \Omega, \tau] \stackrel{ind}{\sim}$

$$N_p((\tau_i X_i^T X_i + \Omega)^{-1}(\tau_i X_i^T Y_i + \Omega Z_i \gamma), (\tau_i X_i^T X_i + \Omega)^{-1})$$

(ii) $[\gamma | Y, \beta, \Omega, \tau] \sim$

$$N_q((\sum_{i=1}^m Z_i^T \Omega Z_i + D^{-1})^{-1}(\sum_{i=1}^m Z_i^T \Omega \beta_i), (\sum_{i=1}^m Z_i^T \Omega Z_i + D^{-1})^{-1})$$

(iii) $[\Omega | Y, \beta, \gamma, \tau] \sim W_p(\alpha + m, R + \frac{1}{2} \sum_{i=1}^m (\beta_i - Z_i \gamma)(\beta_i - Z_i \gamma)^T)$

(iv) For $i = 1, \dots, m$,

$$[\tau_i | Y, \beta, \gamma, \Omega] \sim G(a_i + \frac{1}{2}n_i, b_i + \frac{1}{2}(Y_i - X_i \beta_i)^T(Y_i - X_i \beta_i))$$

For Model 2, the empirical estimators of the posterior means of β_i and μ_i have the same forms (4.5) and (4.7). The Rao-Blackwellized estimators $\hat{\beta}_i^{(RB)}$ and $\hat{\mu}_i^{(RB)}$ are the estimators given by (4.6) and (4.8) with $\tau_e^{(k)}$ replaced by $\tau_i^{(k)}$. Estimator of posterior variance is $\hat{V}(\mu_i)$ given by (4.11) with $\tau_e^{(k)}$ replaced by $\tau_i^{(k)}$.

The Rao-Blackwellized estimators of the posterior mean and posterior variance of σ_i^2 under Model 2 are:

$$\hat{\sigma}_i^{2(RB)} = \frac{1}{G} \sum_{k=1}^G [b_i + \frac{1}{2}(Y_i - X_i\beta_i^{(k)})^T(Y_i - X_i\beta_i^{(k)})](a_i + \frac{1}{2}n_i - 1)^{-1} \quad (4.17)$$

and

$$\begin{aligned} \hat{V}(\sigma_i^2) &= \frac{1}{G} \sum_{k=1}^G [b_i + \frac{1}{2}(Y_i - X_i\beta_i^{(k)})^T(Y_i - X_i\beta_i^{(k)})]^2 \\ &\quad \times (a_i + \frac{1}{2}n_i - 1)^{-2} [(a_i + \frac{1}{2}n_i - 2)^{-1} + 1] \\ &\quad - \frac{1}{G^2} [\sum_{k=1}^G (b_i + \frac{1}{2}(Y_i - X_i\beta_i^{(k)})^T(Y_i - X_i\beta_i^{(k)})) \\ &\quad \times (a_i + \frac{1}{2}n_i - 1)^{-1}]^2 \end{aligned} \quad (4.18)$$

4.2.3 Random error variance model

In Model 2, we assumed unequal error variances for the sampling errors. Kleffe and Rao (1992) used a simple random error variance model to derive the best linear unbiased predictors for small area means. In this section we extend their model to the multi-level case. We assume random effect models on both regression coefficients β_i and sampling error variances σ_i^2 , which leads to Model 3 given below.

Model 3:

- (i) Same as in Model 2;
- (ii) Same as in Model 2;
- (iii) Conditional on η and λ , τ_i 's are independent with

$$\tau_i | \eta, \lambda \stackrel{iid}{\sim} G(\eta, \lambda), \quad (4.19)$$

where $\tau_i = \sigma_i^{-2}$;

- (iv) Marginal prior distributions: $\gamma \sim N_q(0, D)$, $\Omega \sim W_p(\alpha, R)$, $\eta \sim U^+$ and $\lambda \sim U^+$, where U^+ denotes a uniform distribution over a subset of R^+ with large but finite length, D , α and R are known.

Remark 3.1: In Model 3, we assume that τ_i 's are iid gamma random variables with unknown hyperparameters η and λ . Thus we have population models for both regression coefficient β_i and sampling variance σ_i^2 . In Model 1 and Model 2, we considered modelling β_i only and assumed vague proper prior distributions on σ_i^2 or σ_i^2 .

Remark 3.2: Assumption (iii) may not be a good population model for all τ_i 's. Alternatively, we can model τ_i in a more realistic way, as in the case of β_i , by specifying a regression model for the logarithm of τ_i . This may require some auxiliary information related to τ_i . In the data analysis, however, we simply used $G(\eta, \lambda)$ as the population model for τ_i .

The full conditional distributions for Gibbs sampling under Model 3 are given by:

- (i) For $i = 1, \dots, m$, $[\beta_i | Y, \tau, \gamma, \Omega, \eta, \lambda] \stackrel{ind}{\sim}$

$$N_p((\tau_i X_i^T X_i + \Omega)^{-1}(\tau_i X_i^T Y_i + \Omega Z_i \gamma), (\tau_i X_i^T X_i + \Omega)^{-1})$$

- (ii) For $i = 1, \dots, m$,

$$[\tau_i | Y, \beta, \gamma, \Omega, \eta, \lambda] \stackrel{ind}{\sim} G(\eta + \frac{m}{2}, \frac{1}{2}(Y_i - X_i \beta_i)^T (Y_i - X_i \beta_i) + \lambda)$$

- (iii) $[\gamma | Y, \beta, \tau, \eta, \lambda] \sim$

$$N_q((\sum_{i=1}^m Z_i^T \Omega Z_i + D^{-1})^{-1}(\sum_{i=1}^m Z_i^T \Omega \beta_i), (\sum_{i=1}^m Z_i^T \Omega Z_i + D^{-1})^{-1})$$

$$(iv) \quad [\Omega|Y, \beta, \sigma^2, \gamma, \eta, \lambda] \sim W_p(\alpha + m, R + \frac{1}{2} \sum_{i=1}^m (\beta_i - Z_i \gamma)(\beta_i - Z_i \gamma)^T)$$

$$(v) \quad [\eta|Y, \beta, \tau, \gamma, \Omega, \lambda] \propto [\Gamma(\eta)]^{-m} \lambda^{m\eta} (\prod_{i=1}^m \tau_i)^\eta$$

$$(vi) \quad [\lambda|Y, \beta, \tau, \gamma, \Omega, \eta] \sim G(m\eta + 1, \sum_{i=1}^m \tau_i)$$

For Model 3, the posterior estimators of β_i and μ_i have the same forms as those given for Model 2. However, the Rao-Blackwellized estimators of the posterior mean and posterior variance of σ_i^2 are different. Under Model 3, they are given by

$$\hat{\sigma}_i^{2(RB)} = \frac{1}{G} \sum_{k=1}^G [\lambda^{(k)} + \frac{1}{2} (Y_i - X_i \beta_i^{(k)})^T (Y_i - X_i \beta_i^{(k)})] (\eta^{(k)} + \frac{1}{2} n_i - 1)^{-1} \quad (4.20)$$

and

$$\begin{aligned} \hat{V}(\sigma_i^2) &= \frac{1}{G} \sum_{k=1}^G [\lambda^{(k)} + \frac{1}{2} (Y_i - X_i \beta_i^{(k)})^T (Y_i - X_i \beta_i^{(k)})]^2 \\ &\quad \times (\eta^{(k)} + \frac{1}{2} n_i - 1)^{-2} [(\eta^{(k)} + \frac{1}{2} n_i - 2)^{-1} + 1] \\ &\quad - \frac{1}{G^2} [\sum_{k=1}^G (\lambda^{(k)} + \frac{1}{2} (Y_i - X_i \beta_i^{(k)})^T (Y_i - X_i \beta_i^{(k)})) \\ &\quad \times (\eta^{(k)} + \frac{1}{2} n_i - 1)^{-1}]^2 \end{aligned} \quad (4.21)$$

For the sampling issues, $[\eta|Y, \beta, \tau, \gamma, \Omega, \lambda]$ is known only up to a multiplicative constant. However, since $[\eta|Y, \beta, \tau, \gamma, \Omega, \lambda]$ is a log-concave function of η (Arora and Lahiri, 1995), adaptive rejection sampling method (Gilks et al., 1995) can be used.

4.3 Data Analysis

4.3.1 Small area estimation

Following Holt and Moura (1993), we consider the estimation of household income in some counties (small areas) of Brazil. Our data set contains 10 small areas with

28 observations in each area. However, as noted in Holt and Moura (1993), “It must be stressed that this data set is far from ideal insofar as the underlying model assumptions are concerned.” Therefore, the analysis presented here should be viewed only as an illustration of the fully Bayesian inference procedure.

Let y_{ij} denote the j th household’s income in the i th small area. There are two unit level auxiliary variables, namely x_1 and x_2 , where x_1 denotes the number of rooms in a household and x_2 denotes the educational attainment of Head of Household. The sampling model is given by

$$\begin{aligned} y_{ij} &= x_{ij}^T \beta_i + e_{ij} \\ &= \beta_{0i} + x_{1ij} \beta_{1i} + x_{2ij} \beta_{2i} + e_{ij} \end{aligned} \quad (4.22)$$

where x_{1ij} denotes the number of rooms in the j th household of small area i and x_{2ij} denotes the corresponding educational attainment of Head of Household. Values of x_{1ij} and x_{2ij} are centered around their respective overall sample means and e_{ij} is the sampling error variable with its distribution specified by the three error variance models discussed in Section 2.

In the sampling model (4.22), β_i is the random regression coefficient corresponding to the i th small area and is modelled as

$$\beta_{0i} = \gamma_0 + v_{0i} \quad (4.23)$$

$$\beta_{1i} = \gamma_{10} + \gamma_{11} z_i + v_{1i} \quad (4.24)$$

$$\beta_{2i} = \gamma_{20} + \gamma_{21} z_i + v_{2i} \quad (4.25)$$

where $\gamma = (\gamma_0, \gamma_{10}, \gamma_{11}, \gamma_{20}, \gamma_{21})^T$ is the unknown vector of fixed regression parameters, $v_i = (v_{0i}, v_{1i}, v_{2i})^T$ is the i th small area random effect vector distributed as

$v_i \sim N_3(0, \Phi)$, and z_i is an area level variable defined as the average number of cars per household in each small area. Value of z_i is also centered around its overall sample mean.

The i th small area mean is $\mu_i = \bar{X}_i^T \beta_i = \beta_{0i} + \bar{X}_{1i} \beta_{1i} + \bar{X}_{2i} \beta_{2i}$, where $\bar{X}_{1i}$ and $\bar{X}_{2i}$ are the i th small area population means of the auxiliary variables x_1 and x_2 , respectively. Table 4.1 gives the summary statistics that we need in our computation and discussion. To estimate μ_i , we first obtain the posterior estimate $\hat{\beta}_i$ of β_i , then estimate μ_i as $\hat{\mu}_i = \bar{X}_i^T \hat{\beta}_i$.

We use the three models discussed in Section 2 for our data analysis. Vague proper priors on unknown parameters are specified as follows: $\gamma \sim N_5(0, D)$ where $D = \text{diag}(10^4, 10^4, 10^4, 10^4, 10^4)$, thus $\gamma_0, \gamma_{10}, \gamma_{11}, \gamma_{20}, \gamma_{21}$ are assumed to be independent normal variables with a mean of 0 and a standard deviation of 100, so that a 95% prior interval is around ± 200 . A Wishart prior $W_3(\alpha, R)$ is specified for the inverse covariance matrix $\Omega = \Phi^{-1}$. To represent vague prior knowledge, we have chosen the degrees of freedom α for this distribution to be as small as possible, i.e., $\alpha = 3$,

the rank of Ω . The scale matrix R is specified as $R = \begin{pmatrix} 1 & 0.001 & 0.001 \\ 0.001 & 1 & 0.001 \\ 0.001 & 0.001 & 1 \end{pmatrix}$,

which represents our prior guess at the order of magnitude of the covariance matrix.

For Model 1 and Model 2, a gamma prior $G(0.001, 0.001)$ is assumed for τ_e and τ_i 's.

For Model 3, $\tau_i \sim G(\eta, \lambda)$, and η and λ are assumed to be independently distributed as $U(0, 10000)$, i.e., the uniform distribution over a large interval.

We implemented the Gibbs sampler for the three models using the BUGS program (Spiegelhalter et al., 1995), aided by CODA Splus function (Best et al., 1995)

for assessing convergence. The advantage of BUGS program is that it frees us from computational details. After a “burn-in” period of 2000 iterations, 5000 iterations were kept for analysis.

Table 4.2 presents the estimated posterior small area means $\hat{\mu}_i^{(E)}$ and $\hat{\mu}_i^{(RB)}$ for the three models. From the results in Table 4.2, first we note that for each model, $\hat{\mu}_i^{(E)}$ and $\hat{\mu}_i^{(RB)}$ give almost the same point estimates. Between the three models, there is not much difference among the estimated posterior small area means, which indicates that the posterior mean is not sensitive to the choice of sampling error variances. The estimated posterior means for Model 3 are in between those for Model 1 and Model 2. The results for Model 3 are closer to the results for Model 1.

To compare the performance of $\hat{\mu}_i^{(E)}$ and $\hat{\mu}_i^{(RB)}$, we also calculated the simulation standard errors of $\hat{\mu}_i^{(E)}$ and $\hat{\mu}_i^{(RB)}$, which are respectively the sample standard errors of $\{\bar{X}_i^T \hat{\beta}_i^{(E)}\}$ and $\{\bar{X}_i^T E[\beta_i|Y, \gamma^{(k)}, \Omega^{(k)}, \tau_e^{(k)}]\}$. Table 4.3 presents the simulation standard errors. It is clear from Table 4.3 that the Rao-Blackwellized estimator $\hat{\mu}_i^{(RB)}$ has much smaller simulation standard error than the empirical estimator $\hat{\mu}_i^{(E)}$ for all the models. In all cases the standard error of $\hat{\mu}_i^{(RB)}$ is about 50% to 75% of the standard error of $\hat{\mu}_i^{(E)}$. The benefit of “Rao-Blackwellizing” is obvious. Thus $\hat{\mu}_i^{(RB)}$ is more stable than $\hat{\mu}_i^{(E)}$ when used to produce point estimates for the posterior means in computational Bayesian analysis.

To compare the posterior variances of μ_i , Table 4.4 gives the Rao-Blackwellized estimator $\hat{V}(\mu_i)$ for the three models. We note from Table 4.4 that Model 2 leads to the smallest posterior variance of μ_i among the three models for all areas except area 5, which has the largest estimated posterior mean among the ten areas. In

Table 4.1, the values of y , x , and z for area 5 are substantially larger than those values for the other areas, which suggests that area 5 is significantly different from the other areas and could be considered as an outlier area. Model 1 has the largest estimated posterior variances. For a better understanding of the results of the estimated posterior means and variances, it is necessary to look at the estimated error variances under each model, since our models are based on different assumptions on error variances. Table 4.5 shows for each small area the Rao-Blackwellized estimates $\hat{\sigma}_i^{2(RB)}$ of the sampling error variances under the three models. For comparison, we also calculated the ordinary least square (OLS) estimates of σ_i^2 using the unit-level data only.

First we note from the OLS estimates that there are large variations among the ten areas. Model 1 assumes an equal error variance σ_e^2 for all areas and σ_e^2 is estimated by $\hat{\sigma}_e^{2(RB)} = 76.86$, which is much smaller than the OLS estimates for some areas. Model 2 assumes unequal error variance σ_i^2 across areas. Under Model 2, the estimated error variances $\hat{\sigma}_i^{2(RB)}$ to some extent show the feature of the areas; $\hat{\sigma}_i^{2(RB)}$ are consistent with the trend of the OLS estimates. The most notable result is $\hat{\sigma}_5^{2(RB)} = 121.65$ and $\hat{\sigma}_8^{2(RB)} = 159.94$, which show that there are larger variations within small areas 5 and 8.

Model 3 assumes σ_i^2 's to be random variables distributed as $G(\eta, \lambda)$. Under Model 3, all $\hat{\sigma}_i^{2(RB)}$ tend to be equal to and have moved toward $\hat{\sigma}_e^{2(RB)} = 76.86$. This is the reason that in Table 4.2 the estimated posterior small area means under Model 3 are similar to those under Model 1. However Model 3 is different from Model 1 since Model 3 allows different σ_i^2 , which leads to smaller estimated posterior variance

under Model 3 in Table 4.4 than under Model 1 for all areas except areas 5 and 8. The results for Model 3 are in between the results for Model 1 and Model 2. Model 2 assumes unequal error variances without any restrictions on σ_i^2 's. Model 3 assumes a gamma distribution, $G(\eta, \lambda)$, on σ_i^2 . The results in Table 4.2 and Table 4.4 indicate that $G(\eta, \lambda)$ is not a good model for σ_i^2 's for the data set used here.

4.3.2 Model selection and comparison

Model selection

In order to study how the data support each model, we calculated the cross-validation predictive densities for each data point y_{ij} , i.e., the conditional density $f(y_{ij}|Y_{(ij)})$, where $Y_{(ij)}$ denotes all data except y_{ij} . $f(y_{ij}|Y_{(ij)})$ is attractive in that it suggests which values of y_{ij} are likely when the model is fitted to all of the observations except y_{ij} (Gelfand, 1995). We looked at the value of $f(y_{ij}|Y_{(ij)})$ at the observed data point, i.e., the CPO value

$$CPO_{ij} = f(y_{ij,obs}|Y_{(ij),obs})$$

where $y_{ij,obs}$ denotes the observed data point. Since CPOs are nothing but the observed likelihoods, larger CPOs suggest a more likely model. For our three models, we calculated the CPOs for all data points using the output from the Gibbs sampler, as described in Chapter 2. We present a CPO plot for the three models in Figure 4.1. Clearly Model 2 is the best model, and Model 3 is slightly better than Model 1. Also there are many small CPO values for all the three models, which indicates that our model assumptions are not well-satisfied by the data set. More works need to be done in getting better fitting model, especially for areas 5 and 8.

A simulation study

From the small area estimation analysis and model selection, Model 2 is the best model for the data set used here. It would be interesting to know if Model 2, i.e., the unequal error variance model, can be used to fit the data sets generated from Model 1 and Model 3. For this purpose, we did a model selection simulation study. At first, Model 1 was used as the true model to generate a data set, we then fit Model 1, Model 2 and Model 3 to this data set and computed the CPO values under each model. Since Model 1 can be regarded as a special case of Model 2, we would expect that the CPO values will support both Model 1 and Model 2 in this case. We also used Model 2 and Model 3 as the true model respectively to generate a data set and apply the same model fitting procedure to calculate the CPO values under each model.

Since our interest is on the assumption of sampling error variances, we will use the following simplified Model 1, Model 2 and Model 3 as the true models in our simulation study.

Model 1 (equal error variance model):

- $y_{ij} \sim N(\mu_i, \sigma_e^2), (i = 1, \dots, 10; j = 1, \dots, 28),$
- $\mu_i \sim N(11.65, 2.55), (i = 1, \dots, 10),$

where $\sigma_e^2 = 76.86$ is obtained from the estimated error variance under Model 1 in Table 4.5. Further, 11.65 and 2.55 were estimated from $\hat{\mu}_i^{(RB)}$ under Model 1 in Table 4.2.

Model 2 (unequal error variance model):

- $y_{ij} \sim N(\mu_i, \sigma_i^2)$, ($i = 1, \dots, 10; j = 1, \dots, 28$),
- $\mu_i \sim N(11.31, 2.51)$, ($i = 1, \dots, 10$),

where σ_i^2 are the estimated error variances under Model 2 given in Table 4.5. Further, 11.31 and 2.51 were estimated from $\hat{\mu}_i^{(RB)}$ under Model 2 in Table 4.2.

Model 3 (random error variance model):

- $y_{ij} \sim N(\mu_i, \sigma_i^2)$, ($i = 1, \dots, 10; j = 1, \dots, 28$),
- $\sigma_i^2 \sim IG(58, 4324)$, ($i = 1, \dots, 10$),
- $\mu_i \sim N(11.52, 2.52)$, ($i = 1, \dots, 10$),

where IG denotes inverse gamma distribution, parameter values 58 and 4324 were obtained under Model 3 in the data analysis. Further, 11.52 and 2.52 were estimated from $\hat{\mu}_i^{(RB)}$ under Model 3 in Table 4.2.

Table 4.6 shows the model fitting results from the simulation study. The summary statistic used here is the log of the product of the CPO values for each data point. When the true model is Model 1, the $\log(\text{CPO})$ values for fitted Model 1 and Model 2 are about the same, whereas Model 3 has smaller $\log(\text{CPO})$ value than Models 1 and 2. Thus, when the sampling error variance is the same across small areas, random error variance model is not appropriate, but unequal error variance model can be used. The result is reasonable since Model 1 is a special case of Model 2. However, if the true model is Model 2, Model 2 itself is the most appropriate model and Model 1 is the worst fitted model with $\log(\text{CPO})$ equal to -732.78. Model 3 is in between Model 2 and Model 1. If the true model is Model 3, then Model 3 itself is the best fitted model and Model 1 is the worst. The simulation result suggests

that a random error variance model (Model 3) is different from a non-random error variance model (Model 1 and Model 2). It is important in practice to distinguish between unequal but fixed variances and random error variances.

4.3.3 Sensitivity analysis

According to the data analysis and model selection in this section, the unequal error variance model, i.e., Model 2 appears to be the best fitted model. In Model 2, the error variances $\tau_i = \sigma_i^{-2}$ are assumed to be independent with prior distributions $G(a_i, b_i)$, or σ_i^2 with the inverse gamma $IG(a_i, b_i)$, where a_i and b_i are known values chosen to reflect our prior knowledge about σ_i^2 . As the number of small areas m increases, the number of variance components σ_i^2 will increase. The posterior estimates of μ_i and σ_i^2 should not be sensitive to the choice of prior informations on σ_i^2 and the posterior distribution should be proper. In our data analysis, a_i and b_i were chosen to be 0.001. Thus we used proper priors with very small parameter values for the variance components to reflect our vague knowledge about σ_i^2 . Since all our priors are proper, we avoided the problem of improper posteriors (Hobert and Casella, 1996). In order to test the sensitivity of the posterior estimates on the choice of a_i and b_i under Model 2, we set a_i and b_i at six different values, i.e., 0.0001, 0.001, 0.01, 0.1, 1, and 10, where $IG(0.0001, 0.0001)$, $IG(0.001, 0.001)$, and $IG(0.01, 0.01)$ can be viewed as approximate noninformative priors whereas $IG(1, 1)$ and $IG(10, 10)$ are informative priors. Table 4.7 presents the posterior estimates for the small area means under Model 2 using the different priors on σ_i^2 .

From the result shown in Table 4.7, the small area mean estimates are very stable: they are not sensitive to the choice of priors on σ_i^2 . This is actually due to

the fact that

$$E(\beta_i|Y, \gamma, \Omega, \tau) = (\tau_i X_i^T X_i + \Omega)^{-1}(\tau_i X_i^T Y_i + \Omega Z_i \gamma), \quad (4.26)$$

where $\tau_i = \sigma_i^{-2}$. Intuitively, the cancellation of the effects of σ_i^2 in the numerator and denominator of (4.26) leads to the insensitivity of posterior mean to the variation of σ_i^2 . Table 4.8 presents the posterior estimates for σ_i^2 using the different priors on σ_i^2 .

As we can see from Table 4.8, when a_i and b_i are small (< 0.01), there is no difference among the estimates. As a_i and b_i increase, the estimates $\hat{\sigma}_i^{2(RB)}$ become smaller. Since

$$[\tau_i|Y, \beta, \gamma, \Omega] \sim G(a_i + \frac{1}{2}n_i, b_i + \frac{1}{2}(Y_i - X_i\beta_i)^T(Y_i - X_i\beta_i)), \quad (4.27)$$

when a_i and b_i are small, we have approximate noninformative priors, the sample effects $\frac{1}{2}n_i$ and $\frac{1}{2}(Y_i - X_i\beta_i)^T(Y_i - X_i\beta_i)$ dominate the prior information a_i and b_i . However, if there is strong prior information on a_i and b_i , for example, $a_i = b_i = 10$, then the posterior estimates of σ_i^2 will be significantly different from the ones under noninformative priors. So $\hat{\sigma}_i^{2(RB)}$ is sensitive to the prior choice, but it is fairly stable when the prior is approximately noninformative.

4.4 Summary

In this Chapter, we have presented hierarchical Bayes methods for small area estimation using multi-level models. Clearly it is not easy to provide a suitable model for all small areas with satisfactory results, even if the Markov chain Monte Carlo Bayesian methods such as Gibbs sampling enable us to fit the data using Bayesian

models of virtually unlimited complexity. The size and homogeneity of the areas and the availability of good auxiliary information will affect the final results. Models which prove suitable in some situations may be unsuitable in others. Nevertheless, the general hierarchical Bayes methodology is applicable to a wide variety of situations for estimation of small area parameters. Model selection and choice is an important part of the hierarchical Bayes analysis. It is also important to compare the hierarchical Bayes method with other widely used methods in small area estimation such as empirical Bayes and the EBLUP method.

Small Area	Sample Mean			Area z_i	Population Mean	
	$\bar{y}_i$	$\bar{x}_{1i}$	$\bar{x}_{2i}$		$\bar{X}_{1i}$	$\bar{X}_{2i}$
1	11.082	0.226	0.270	0.454	0.062	0.206
2	7.900	0.940	-0.158	0.644	0.684	0.240
3	13.482	1.083	0.627	0.686	1.039	0.521
4	6.532	0.726	-0.266	0.742	1.156	0.344
5	19.514	2.155	1.056	0.982	2.237	0.826
6	11.205	1.119	0.342	0.555	0.622	0.118
7	8.717	0.083	-0.016	0.634	0.440	0.038
8	12.804	0.190	-0.301	0.622	0.301	0.026
9	10.179	0.512	0.056	0.739	0.901	0.369
10	10.004	0.333	-0.051	0.535	0.280	0.057

Table 4.1: Summary of Small Area Data

Small Area	Model 1		Model 2		Model 3	
	$\hat{\mu}_i^{(E)}$	$\hat{\mu}_i^{(RB)}$	$\hat{\mu}_i^{(E)}$	$\hat{\mu}_i^{(RB)}$	$\hat{\mu}_i^{(E)}$	$\hat{\mu}_i^{(RB)}$
1	10.30	10.30	10.22	10.23	10.28	10.27
2	10.37	10.35	9.84	9.84	10.23	10.24
3	13.11	13.08	13.02	13.00	13.08	13.07
4	11.42	11.43	10.95	10.95	11.27	11.26
5	18.36	18.35	17.88	17.87	18.15	18.14
6	10.45	10.47	10.21	10.21	10.38	10.37
7	9.81	9.80	9.59	9.58	9.74	9.73
8	11.09	11.09	10.29	10.30	10.77	10.77
9	11.73	11.71	11.32	11.34	11.58	11.56
10	9.93	9.94	9.80	9.79	9.88	9.87

Table 4.2: Estimated Posterior Small Area Means

Small Area	Model 1		Model 2		Model 3	
	$\hat{\mu}_i^{(E)}$	$\hat{\mu}_i^{(RB)}$	$\hat{\mu}_i^{(E)}$	$\hat{\mu}_i^{(RB)}$	$\hat{\mu}_i^{(E)}$	$\hat{\mu}_i^{(RB)}$
1	0.978	0.604	0.817	0.506	0.912	0.578
2	1.061	0.612	0.862	0.498	1.014	0.615
3	1.092	0.609	1.090	0.548	1.091	0.605
4	1.260	0.692	1.101	0.604	1.195	0.695
5	1.286	0.630	1.583	0.878	1.341	0.716
6	0.956	0.523	0.930	0.481	0.937	0.531
7	0.989	0.520	0.978	0.480	0.994	0.529
8	1.199	0.904	1.208	0.842	1.219	0.924
9	1.106	0.595	0.997	0.524	1.083	0.600
10	0.961	0.572	0.869	0.513	0.911	0.556

Table 4.3: Simulation Standard Errors

Small Area	Model 1 $\hat{V}(\mu_i)$	Model 2 $\hat{V}(\mu_i)$	Model 3 $\hat{V}(\mu_i)$
1	0.942	0.658	0.826
2	1.134	0.724	1.022
3	1.172	1.167	1.170
4	1.511	1.220	1.416
5	1.634	2.455	1.845
6	0.924	0.870	0.894
7	0.966	0.933	0.938
8	1.434	1.417	1.470
9	1.206	1.014	1.136
10	0.912	0.760	0.829

Table 4.4: Estimated Posterior Variance of μ_i

Small Area	Estimated Sampling Error Variances			
	OLS	Model 1	Model 2	Model 3
1	38.17	76.86 (6.78)	40.18 (12.02)	63.60 (13.33)
2	31.75	76.86 (6.78)	34.24 (10.37)	62.13 (13.79)
3	81.26	76.86 (6.78)	94.77 (28.70)	79.58 (14.60)
4	48.73	76.86 (6.78)	52.01 (15.51)	67.27 (12.77)
5	115.98	76.86 (6.78)	121.65 (36.81)	87.70 (18.09)
6	90.74	76.86 (6.78)	94.35 (27.75)	79.78 (14.36)
7	101.67	76.86 (6.78)	101.67 (29.62)	82.14 (15.19)
8	135.65	76.86 (6.78)	159.94 (48.34)	97.96 (23.98)
9	59.10	76.86 (6.78)	63.37 (18.99)	70.57 (12.63)
10	62.86	76.86 (6.78)	65.72 (19.47)	71.22 (12.62)

Table 4.5: Estimated Sampling Error Variances

Fitted Model	True Model		
	Model 1	Model 2	Model 3
Model 1	-713.32	-732.78	-736.04
Model 2	-713.67	-702.75	-726.70
Model 3	-724.41	-714.42	-715.85

Table 4.6: Log of the Product of the CPO Values

Small Area	$IG(a_i, b_i)$					
	0.0001	0.001	0.01	0.1	1	10
1	10.23	10.23	10.23	10.24	10.25	10.37
2	9.84	9.84	9.84	9.83	9.82	9.62
3	13.00	13.00	13.01	13.01	13.07	13.09
4	10.95	10.95	10.95	10.95	10.94	10.61
5	17.86	17.87	17.85	17.76	17.78	18.28
6	10.21	10.21	10.21	10.21	10.25	10.28
7	9.58	9.58	9.59	9.58	9.63	9.57
8	10.29	10.30	10.30	10.26	10.37	10.86
9	11.34	11.34	11.35	11.32	11.32	11.23
10	9.79	9.79	9.80	9.79	9.82	9.92

Table 4.7: Comparison of Estimated Small Area Means

Small Area	$IG(a_i, b_i)$					
	0.0001	0.001	0.01	0.1	1	10
1	40.09	40.10	40.05	39.64	37.14	22.29
2	34.19	34.18	34.17	33.97	31.74	19.05
3	94.48	94.49	94.42	93.76	86.73	50.60
4	52.08	52.08	52.04	51.63	48.21	28.82
5	121.60	121.70	121.60	121.40	113.70	66.75
6	94.03	94.03	93.83	92.96	87.21	52.90
7	102.30	102.30	102.20	101.40	94.85	57.58
8	160.10	160.00	159.90	159.10	147.60	86.61
9	63.46	63.46	63.38	62.99	58.46	34.85
10	65.88	65.87	65.89	65.40	60.76	36.60

Table 4.8: Comparison of Estimated Sampling Error Variances

Figure 4.1: Multi-level Models CPO Plot

Chapter 5

Pseudo Hierarchical Bayes Small Area Estimation Using Sampling Weights

5.1 Introduction

Unit-level random effects models including nested error regression models and multi-level models are often used in small area estimation to obtain efficient model-based estimators of small area means. Such estimators typically do not make use of the survey weights (Battese, Harter and Fuller 1988; Prasad and Rao 1990; You and Rao, 1997). As a result, the estimators are not design consistent unless the sampling design is self-weighting within areas. On the other hand, direct design-based estimators are design consistent but fail to borrow strength from the related small areas. Kott (1989) advocated the use of design-consistent model-based estimators (i.e., model assisted estimators). He derived a design-consistent estimator of a small area mean under a simple random effects model. He also proposed an estimator of its mean squared error (MSE) which is model unbiased as well as design-consistent.

But the MSE estimator is quite unstable and can even take negative values. Prasad and Rao (1999) proposed a two-step procedure to obtain a pseudo EBLUP estimator of a small area mean. In the first step, an area-level model is derived from the unit-level model using survey weights and optimal small area estimates are obtained assuming the variance parameters in the model are known. In the second step, the variance parameters are estimated from the unit-level model and then used in the optimal first-step estimator. The pseudo EBLUP estimators are design-consistent and borrow strength across small areas. Prasad and Rao (1999) also derived a stable model-based MSE estimator, assuming the number of small areas is large.

In this Chapter, a two-step hierarchical Bayes approach is used to obtain posterior estimators of the small area means. In the first step, the Bayes estimator (posterior mean) and a measure of its variability (posterior variance) are derived from the aggregated area-level model, assuming the variance parameters are known. In the second step, the posterior distribution of variance parameters from the unit-level model is employed to derive a pseudo posterior mean which is used as the estimate of a small area mean and a pseudo posterior variance which is used as a measure of variability of the estimate. In section 4 a limited simulation study shows that the pseudo posterior estimators are efficient. Our results suggest that the proposed approach is very promising and attractive. In section 5, the pseudo hierarchical Bayes approach is extended to general multi-level models.

5.2 Simple Random Effects Model

A simple random effects unit level full model is given by

$$y_{ij} = \theta_i + e_{ij}, j = 1, \dots, N_i; i = 1, \dots, m, \quad (5.1)$$

where y_{ij} are the unit population values, e_{ij} are uncorrelated random errors with $E(e_{ij}) = 0$ and $V(e_{ij}) = \sigma_i^2$, and N_i is the population size in the i -th area ($i = 1, \dots, m$). For simplicity, we take θ_i as the small area mean $\bar{Y}_i = \sum_j y_{ij}/N_i$. Note that $\bar{Y}_i = \theta_i + \bar{E}_i$ and $\bar{E}_i = \sum_j e_{ij}/N_i \approx 0$ if N_i is large, so that $\bar{Y}_i \approx \theta_i$. It is customary to assume equal error variances $\sigma_i^2 = \sigma_e^2$, although the case of unequal and random error variances has also been studied (Kleffe and Rao, 1992; and Arora and Lahiri, 1997; You and Rao, 1997).

The linking model or population model assumes the small area mean θ_i is a random variable satisfying

$$\theta_i = \mu + v_i, i = 1, \dots, m, \quad (5.2)$$

where the v_i 's are uncorrelated random variables with $E(v_i) = 0$ and $V(v_i) = \sigma_v^2$. Further, v_i and e_{ij} are assumed to be uncorrelated for all i and j .

Assuming that the full model (5.1) also holds for the sample $\{y_{ij}, j = 1, \dots, n_i; i = 1, \dots, m\}$ and combining the sample model with the linking model, we get the unit-level simple random effects model

$$y_{ij} = \mu + v_i + e_{ij}, j = 1, \dots, n_i; i = 1, \dots, m. \quad (5.3)$$

This is the well known two variance components model and is used in small area estimation to obtain model-based estimators of small area means. However, the

model-based estimators ignore the sampling design, particularly, the design weights attached to the sample values $\{y_{ij}, j = 1, \dots, n_i; i = 1, \dots, m.\}$.

To bring the design weights into the estimation, suppose $\tilde{w}_{ij}$ denotes the basic design weight attached to the j -th sample unit ($j = 1, \dots, n_i$) in the i -th small area ($i = 1, \dots, m$). A direct design-based estimator of small area mean θ_i is given by the ratio estimator

$$\bar{y}_{iw} = \frac{\sum_{j=1}^{n_i} \tilde{w}_{ij} y_{ij}}{\sum_{j=1}^{n_i} \tilde{w}_{ij}} = \sum_{j=1}^{n_i} w_{ij} y_{ij}, \quad (5.4)$$

where

$$w_{ij} = \frac{\tilde{w}_{ij}}{\sum_{j=1}^{n_i} \tilde{w}_{ij}}. \quad (5.5)$$

The direct estimator $\bar{y}_{iw}$ is design consistent but fails to borrow strength from other areas. By combining the direct estimator (5.4) with the sampling model (5.3), we get the following aggregated area level model

$$\begin{aligned} \bar{y}_{iw} &= \sum_{j=1}^{n_i} w_{ij} (\mu + v_i + e_{ij}) \\ &= \mu + v_i + \bar{e}_{iw}, \end{aligned} \quad (5.6)$$

where $\bar{e}_{iw} = \sum_{j=1}^{n_i} w_{ij} e_{ij}$ are uncorrelated random variables with $E(\bar{e}_{iw}) = 0$ and $V(\bar{e}_{iw}) = \delta_i = \sigma_e^2 \sum_{j=1}^{n_i} w_{ij}^2$. The aggregated model (5.6) is an area level model similar to the well known Fay-Herriot Model (Fay and Herriot, 1979). We will use the aggregated area level model (5.6) to obtain the posterior estimates of the small area means θ_i .

5.3 Bayesian Analysis

5.3.1 Bayesian estimation

We are interested in obtaining the posterior estimates of the small area mean θ_i based on the design consistent estimator $\bar{y}_{iw}$ using the aggregated area level model (5.6). Under model (5.6) with normal assumption on v_i and e_{ij} , we first obtain the posterior distribution of θ_i given $\bar{Y}_w = \{\bar{y}_{iw}, i = 1, \dots, m\}$, σ_e^2 and σ_v^2 , by assuming a flat prior distribution on μ .

Theorem: If $Y_i|\theta_i \sim N(\theta_i, V_i)$ and $\theta_i|\beta, A \sim N(z_i^T\beta, A)$, where $i = 1, \dots, k$, and usual independence assumptions hold, then by assuming a flat prior on β , the posterior distribution of θ given Y, V and A is:

$$\theta|Y, V, A \sim N((I + A^{-1}VF)^{-1}Y, (I + A^{-1}VF)^{-1}V),$$

where Y and θ are vectors with elements Y_i and θ_i , V is a diagonal matrix with elements V_i and $F = I - Z(Z^T Z)^{-1}Z^T$, $Z = (z_1, \dots, z_k)^T$.

Proof: Straightforward integration with respect to β or using the result of Lindley and Smith (1972).

Matrix Lemma: For any matrices A_1, A_2, C_1 , and C_2 of appropriate dimensions and for which the inverses stated in the result exist, we have

$$C_1^{-1} - C_1^{-1}A_1(A_1^T C_1^{-1}A_1 + C_2^{-1})^{-1}A_1^T C_1^{-1} = (C_1 + A_1 C_2 A_1^T)^{-1}.$$

(Lindley and Smith, 1972)

By using the above theorem and lemma, and after simplification, we obtained

$$\theta_i|\bar{Y}_w, \sigma_e^2, \sigma_v^2 \sim N(\bar{\theta}_i, \bar{V}(\theta_i)), \quad (5.7)$$

where

$$\tilde{\theta}_i = E(\theta_i | \bar{Y}_w, \sigma_e^2, \sigma_v^2) = r_{iw} \bar{y}_{iw} + (1 - r_{iw}) \bar{\mu} \quad (5.8)$$

with

$$r_{iw} = \frac{\sigma_v^2}{\sigma_v^2 + \delta_i}, \quad \bar{\mu} = \left(\sum_{i=1}^m \frac{1}{\sigma_v^2 + \delta_i} \right)^{-1} \left(\sum_{i=1}^m \frac{\bar{y}_{iw}}{\sigma_v^2 + \delta_i} \right).$$

and

$$\begin{aligned} \tilde{V}(\theta_i) = V(\theta_i | \bar{Y}_w, \sigma_e^2, \sigma_v^2) &= r_{iw} \delta_i + (1 - r_{iw})^2 \left(\sum_{i=1}^m \frac{1}{\sigma_v^2 + \delta_i} \right)^{-1} \\ &= g_{1i}(\sigma_e^2, \sigma_v^2) + g_{2i}(\sigma_e^2, \sigma_v^2). \end{aligned} \quad (5.9)$$

Note that the posterior mean $\tilde{\theta}_i$ given by (5.8) is the same as the BLUP estimator of θ_i in Prasad and Rao (1999) and the posterior variance $\tilde{V}(\theta_i)$ given by (5.9) is the same as the mean squared error of the BLUP estimator.

The posterior mean $\tilde{\theta}_i$ and posterior variance $\tilde{V}(\theta_i)$ depend on σ_e^2 and σ_v^2 , whereas σ_e^2 and σ_v^2 are unknown in practice. By taking account of the uncertainty about σ_e^2 and σ_v^2 , we obtain the posterior mean and posterior variance of θ_i given $\bar{Y}_w$ as

$$\begin{aligned} E(\theta_i | \bar{Y}_w) &= E(E(\theta_i | \bar{Y}_w, \sigma_e^2, \sigma_v^2)) \\ &= \int E(\theta_i | \bar{Y}_w, \sigma_e^2, \sigma_v^2) \pi(\sigma_e^2, \sigma_v^2 | \bar{Y}_w) d\sigma_e^2 d\sigma_v^2 \end{aligned} \quad (5.10)$$

and

$$\begin{aligned} V(\theta_i | \bar{Y}_w) &= E(V(\theta_i | \bar{Y}_w, \sigma_e^2, \sigma_v^2)) + V(E(\theta_i | \bar{Y}_w, \sigma_e^2, \sigma_v^2)) \\ &= E_{\sigma_e^2, \sigma_v^2}(\tilde{V}(\theta_i)) + V_{\sigma_e^2, \sigma_v^2}(\tilde{\theta}_i) \\ &= E_{\sigma_e^2, \sigma_v^2}(\tilde{V}(\theta_i)) + E_{\sigma_e^2, \sigma_v^2}(\tilde{\theta}_i^2) - [E_{\sigma_e^2, \sigma_v^2}(\tilde{\theta}_i)]^2 \\ &= \int \tilde{V}(\theta_i) \pi(\sigma_e^2, \sigma_v^2 | \bar{Y}_w) d\sigma_e^2 d\sigma_v^2 \\ &+ \int \tilde{\theta}_i^2 \pi(\sigma_e^2, \sigma_v^2 | \bar{Y}_w) d\sigma_e^2 d\sigma_v^2 - [\int \tilde{\theta}_i \pi(\sigma_e^2, \sigma_v^2 | \bar{Y}_w) d\sigma_e^2 d\sigma_v^2]^2, \end{aligned} \quad (5.11)$$

where $\pi(\sigma_e^2, \sigma_v^2 | \bar{Y}_w)$ is the posterior distribution of σ_e^2 and σ_v^2 given the aggregated data $\bar{Y}_w$, $E_{\sigma_e^2, \sigma_v^2}$ and $V_{\sigma_e^2, \sigma_v^2}$ respectively denote the expectation and variance with respect to the posterior distribution of σ_e^2 and σ_v^2 given $\bar{Y}_w$.

The marginal distribution of $\bar{y}_{iw}$ from the aggregated area-level model (5.6) is

$$\bar{y}_{iw} \sim N(\mu, \sigma_v^2 + \sigma_e^2 \sum_{j=1}^{n_i} w_{ij}^2), \quad i = 1, \dots, m. \quad (5.12)$$

Let $d_i = \sum_{j=1}^{n_i} w_{ij}^2$. Since both σ_e^2 and σ_v^2 are not known in the marginal model (5.12), σ_e^2 and σ_v^2 are not identifiable in the marginal area-level model (5.12) if all the d_i 's are equal, say, $d_i \equiv d_0$. In general, we cannot get reliable estimates of σ_e^2 and σ_v^2 using only $\{\bar{y}_{iw}\}$ based on (5.12). Using $\pi(\sigma_e^2, \sigma_v^2 | \bar{Y}_w)$ in the posterior estimation (5.10) and (5.11) may result in misleading results, particularly in Monte Carlo evaluation. Bell (1998) also indicated similar problems. Thus unit-level data and model should be used to get reliable posterior estimates of σ_e^2 and σ_v^2 .

5.3.2 Pseudo Bayes estimation

By formulating the unit level sampling model (5.3) in a hierarchical Bayes framework, we can obtain the posterior distribution of σ_e^2 and σ_v^2 given the original sample $Y = \{y_{ij}, j = 1, \dots, n_i; i = 1, \dots, m\}$ as $\pi(\sigma_e^2, \sigma_v^2 | Y)$. Replacing $\pi(\sigma_e^2, \sigma_v^2 | \bar{Y}_w)$ by $\pi(\sigma_e^2, \sigma_v^2 | Y)$ in equation (5.10) and (5.11), we obtain the pseudo posterior mean and variance of θ_i denoted as

$$E^p(\theta_i | \bar{Y}_w, Y) = \int E(\theta_i | \bar{Y}_w, \sigma_e^2, \sigma_v^2) \pi(\sigma_e^2, \sigma_v^2 | Y) d\sigma_e^2 d\sigma_v^2 \quad (5.13)$$

and

$$V^p(\theta_i | \bar{Y}_w, Y) = \int \tilde{V}(\theta_i) \pi(\sigma_e^2, \sigma_v^2 | Y) d\sigma_e^2 d\sigma_v^2$$

$$+ \int \bar{\theta}_i^2 \pi(\sigma_e^2, \sigma_v^2 | Y) d\sigma_e^2 d\sigma_v^2 - [\int \bar{\theta}_i \pi(\sigma_e^2, \sigma_v^2 | Y) d\sigma_e^2 d\sigma_v^2]^2, \quad (5.14)$$

where E^p and V^p denote the pseudo posterior expectation and variance, respectively. Numerical evaluation of (5.13) and (5.14) involves two dimensional integration. To obtain $\pi(\sigma_e^2, \sigma_v^2 | Y)$, following Box and Tiao (1973), we first get the joint posterior distribution of $(\mu, \sigma_e^2, \sigma_v^2)$ as

$$\begin{aligned} \pi(\mu, \sigma_e^2, \sigma_v^2 | Y) &\propto (\sigma_e^2)^{-\frac{1}{2} \sum_{i=1}^m (n_i - 1)} \prod_{i=1}^m (\sigma_e^2 + n_i \sigma_v^2)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma_e^2} \sum_{i=1}^m (n_i - 1) S_i^2\right\} \\ &\exp\left\{-\frac{1}{2} \sum_{i=1}^m \frac{n_i}{\sigma_e^2 + n_i \sigma_v^2} (\mu - \bar{y}_i)^2\right\} \pi(\mu) \pi(\sigma_e^2) \pi(\sigma_v^2) \end{aligned} \quad (5.15)$$

where $\bar{y}_i = \frac{1}{n_i} \sum_j y_{ij}$, $S_i^2 = \frac{1}{n_i - 1} \sum_j (y_{ij} - \bar{y}_i)^2$, and $\pi(\mu)$, $\pi(\sigma_e^2)$, $\pi(\sigma_v^2)$ are priors of $(\mu, \sigma_e^2, \sigma_v^2)$. By assuming a flat prior on μ and integrating $\pi(\mu, \sigma_e^2, \sigma_v^2 | Y)$ over μ , we obtained the joint posterior distribution of σ_e^2 and σ_v^2 as

$$\begin{aligned} \pi(\sigma_e^2, \sigma_v^2 | Y) &\propto (\sigma_e^2)^{-\frac{1}{2} \sum_{i=1}^m (n_i - 1)} \prod_{i=1}^m (\sigma_e^2 + n_i \sigma_v^2)^{-\frac{1}{2}} \left(\sum_{i=1}^m \frac{n_i}{\sigma_e^2 + n_i \sigma_v^2}\right)^{-\frac{1}{2}} \\ &\exp\left\{-\frac{1}{2\sigma_e^2} \sum_{i=1}^m (n_i - 1) S_i^2 - \frac{1}{2} \sum_{i=1}^m \frac{n_i \bar{y}_i^2}{\sigma_e^2 + n_i \sigma_v^2}\right\} \\ &\exp\left\{\frac{1}{2} \left(\sum_{i=1}^m \frac{n_i}{\sigma_e^2 + n_i \sigma_v^2}\right)^{-1} \left(\sum_{i=1}^m \frac{n_i \bar{y}_i}{\sigma_e^2 + n_i \sigma_v^2}\right)^2\right\} \pi(\sigma_e^2) \pi(\sigma_v^2). \end{aligned} \quad (5.16)$$

Direct integration of (5.13) and (5.14) involves combining (5.13) and (5.14) with (5.16) and is very complicated, even though we only have two dimensional integration. In general, high dimensional integration may be involved, for example, in the cases of unequal error variances (σ_i^2 unequal) and multi-level models (You and Rao, 1997). Thus Monte Carlo integration should be used and is more feasible in practice. A sampling based approach, in particular, the Gibbs sampling method, is used to estimate the pseudo posterior mean and variance of θ_i .

5.3.3 A Gibbs sampling approach

For Monte Carlo evaluation of (5.13) and (5.14), the Gibbs sampling method (Gelfand and Smith, 1990) is used to draw samples from the joint posterior distribution $\pi(\sigma_e^2, \sigma_v^2 | Y)$. To apply the Gibbs sampling method, we present the unit level sampling model (5.3) in a hierarchical Bayes framework as follows:

$$y_{ij} | \theta_i, \sigma_e^2 \sim N(\theta_i, \sigma_e^2), \quad i = 1, \dots, m, \quad j = 1, \dots, n_i; \quad (5.17)$$

$$\theta_i | \mu, \sigma_v^2 \sim N(\mu, \sigma_v^2), \quad i = 1, \dots, m; \quad (5.18)$$

$$\mu \sim N(0, \tau), \quad \sigma_e^2 \sim IG(a_1, b_1), \quad \sigma_v^2 \sim IG(a_2, b_2); \quad (5.19)$$

where μ has a normal prior, σ_e^2 and σ_v^2 have an inverse gamma (IG) prior distribution, respectively. τ is chosen to be very large and $a_i > 0$, $b_i > 0$, ($i = 1, 2$) are chosen to be very small to reflect our vague knowledge. Thus proper priors are used here to avoid improper posteriors (Hobert and Casella, 1996).

To draw samples from the Gibbs sampler, the following full conditional distributions are needed:

$$\theta_i | Y, \mu, \sigma_e^2, \sigma_v^2 \sim N\left(\left(\frac{n_i}{\sigma_e^2} + \frac{1}{\sigma_v^2}\right)^{-1} \left(\frac{1}{\sigma_e^2} \sum_{j=1}^{n_i} y_{ij} + \frac{\mu}{\sigma_v^2}\right), \left(\frac{n_i}{\sigma_e^2} + \frac{1}{\sigma_v^2}\right)^{-1}\right) \quad (5.20)$$

$$\mu | Y, \theta, \sigma_e^2, \sigma_v^2 \sim N\left(\left(\frac{m}{\sigma_v^2} + \frac{1}{\tau}\right)^{-1} \frac{1}{\sigma_v^2} \sum_{i=1}^m \theta_i, \left(\frac{m}{\sigma_v^2} + \frac{1}{\tau}\right)^{-1}\right) \quad (5.21)$$

$$\sigma_e^2 | Y, \theta, \mu, \sigma_v^2 \sim IG\left(a_1 + \frac{1}{2} \sum_{i=1}^m n_i, b_1 + \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \theta_i)^2\right) \quad (5.22)$$

$$\sigma_v^2 | Y, \theta, \mu, \sigma_e^2 \sim IG\left(a_2 + \frac{m}{2}, b_2 + \frac{1}{2} \sum_{i=1}^m (\theta_i - \mu)^2\right) \quad (5.23)$$

Suppose $\{\theta_i^{(k)}, \mu^{(k)}, \sigma_e^{2(k)}, \sigma_v^{2(k)}; k = 1, \dots, G\}$ are samples generated from the Gibbs sampler based on the Bayesian unit level sampling models (5.17) to (5.19).

Then it follows from (5.13) and (5.14) that the estimated pseudo posterior mean and pseudo posterior variance of θ_i are given respectively by

$$\begin{aligned}\hat{\theta}_i^p &= \frac{1}{G} \sum_{k=1}^G E(\theta_i | \bar{Y}_w, \sigma_e^{2(k)}, \sigma_v^{2(k)}) \\ &= \frac{1}{G} \sum_{k=1}^G \bar{\theta}_i(\sigma_e^{2(k)}, \sigma_v^{2(k)})\end{aligned}\quad (5.24)$$

and

$$\begin{aligned}\hat{V}^p(\theta_i) &= \frac{1}{G} \sum_{k=1}^G [g_{1i}(\sigma_e^{2(k)}, \sigma_v^{2(k)}) + g_{2i}(\sigma_e^{2(k)}, \sigma_v^{2(k)})] \\ &\quad + \frac{1}{G} \sum_{k=1}^G [\bar{\theta}_i(\sigma_e^{2(k)}, \sigma_v^{2(k)})]^2 - [\frac{1}{G} \sum_{k=1}^G \bar{\theta}_i(\sigma_e^{2(k)}, \sigma_v^{2(k)})]^2,\end{aligned}\quad (5.25)$$

where $\bar{\theta}_i$ and g_{1i} , g_{2i} are given by (5.8) and (5.9), respectively.

A naive estimator of $V(\theta_i | \bar{Y}_w)$, $\hat{V}^n(\theta_i)$, is obtained by using only the first component of $\hat{V}^p(\theta_i)$ and ignoring the uncertainty of estimating σ_e^2 and σ_v^2 :

$$\begin{aligned}\hat{V}^n(\theta_i) &= \frac{1}{G} \sum_{k=1}^G V(\theta_i | \bar{Y}_w, \sigma_e^{2(k)}, \sigma_v^{2(k)}) \\ &= \frac{1}{G} \sum_{k=1}^G [g_{1i}(\sigma_e^{2(k)}, \sigma_v^{2(k)}) + g_{2i}(\sigma_e^{2(k)}, \sigma_v^{2(k)})].\end{aligned}\quad (5.26)$$

It is of interest to compute the ratio $\hat{V}^n(\theta_i)/\hat{V}^p(\theta_i)$ to find out the possible under-estimation induced by using $\hat{V}^n(\theta_i)$ only.

By using the output from the Gibbs sampler, we can also find the Rao-Blackwellized estimators of the completely model-based posterior mean $E(\theta_i | Y)$ and posterior variance $V(\theta_i | Y)$ as follows:

$$\begin{aligned}\hat{\theta}_i^m &= \frac{1}{G} \sum_{k=1}^G E(\theta_i | Y, \mu^{(k)}, \sigma_e^{2(k)}, \sigma_v^{2(k)}) \\ &= \frac{1}{G} \sum_{k=1}^G \left(\frac{n_i}{\sigma_e^{2(k)}} + \frac{1}{\sigma_v^{2(k)}} \right)^{-1} \left(\frac{1}{\sigma_e^{2(k)}} \sum_{j=1}^{n_i} y_{ij} + \frac{\mu^{(k)}}{\sigma_v^{2(k)}} \right)\end{aligned}\quad (5.27)$$

and

$$\begin{aligned}\hat{V}^m(\theta_i) &= \frac{1}{G} \sum_{k=1}^G \left[\left(\frac{n_i}{\sigma_e^{2(k)}} + \frac{1}{\sigma_v^{2(k)}} \right)^{-1} + \left(\frac{n_i}{\sigma_e^{2(k)}} + \frac{1}{\sigma_v^{2(k)}} \right)^{-2} \left(\frac{1}{\sigma_e^{2(k)}} \sum_{j=1}^{n_i} y_{ij} + \frac{\mu^{(k)}}{\sigma_v^{2(k)}} \right)^2 \right] \\ &\quad - \frac{1}{G^2} \left[\sum_{k=1}^G \left(\frac{n_i}{\sigma_e^{2(k)}} + \frac{1}{\sigma_v^{2(k)}} \right)^{-1} \left(\frac{1}{\sigma_e^{2(k)}} \sum_{j=1}^{n_i} y_{ij} + \frac{\mu^{(k)}}{\sigma_v^{2(k)}} \right) \right]^2.\end{aligned}\quad (5.28)$$

We will compare the performance of the pseudo posterior mean and pseudo posterior variance with that of the model-based posterior mean and posterior variance in the simulation study.

5.4 A Simulation Study

To evaluate the finite sample performance of the proposed pseudo posterior estimators, we conducted a limited simulation study. We first constructed three fixed finite populations with $m = 30$ small areas in each population. Each area has $N_i = 200$ population units in it. The three synthetic populations were generated from the unit level full models (5.1) and (5.2) by taking $\mu = 50$, $\sigma_e^2 = 225$ and $\sigma_v^2 = 100, 64, 36$, respectively. Thus the ratio σ_v^2/σ_e^2 for Population 1, 2 and 3 is 0.44, 0.28, and 0.16, respectively. After the three synthetic populations were generated, we used the standard ANOVA estimators of σ_e^2 and σ_v^2 , based on all the population values $\{y_{ij}, i = 1, \dots, m; j = 1, \dots, N_i\}$ to obtain the census estimates of σ_e^2 and σ_v^2 . For population 1, the census estimates of σ_e^2 and σ_v^2 are $\bar{\sigma}_e^2 = 221.93$ and $\bar{\sigma}_v^2 = 99.28$, respectively. For population 2, the corresponding census estimates are $\bar{\sigma}_e^2 = 225.47$ and $\bar{\sigma}_v^2 = 63.50$. For population 3, $\bar{\sigma}_e^2 = 226.65$ and $\bar{\sigma}_v^2 = 36.60$. These census estimates will be used in our simulation study for the comparison of pseudo posterior estimators.

From each population, PPS (probability proportional to size) samples within each areas were drawn independently. To implement PPS sampling within each area, size measures x_{ij} for each y_{ij} were generated from an exponential distribution with mean 200. Using these x values, we computed selection probabilities $p_{ij} = x_{ij} / \sum_j x_{ij}$ for each unit y_{ij} and then used them to select PPS with replacement samples. For the three populations, we chose equal sample size $n_i = n$ within each area, by taking $n = 5, 20$ and 40 , respectively.

The basic design weights are given by $\bar{w}_{ij} = n^{-1} p_{ij}^{-1}$ so that $w_{ij} = p_{ij}^{-1} / \sum_j p_{ij}^{-1}$. Using these weights and the associated sample values y_{ij} , we computed the direct ratio estimate $\bar{y}_{iw}$. To compute the pseudo posterior mean estimate $\hat{\theta}_i^p$ and the associated pseudo posterior variance estimates $\hat{V}^p(\theta_i)$, we need to first find the posterior estimates of σ_e^2 and σ_v^2 . Samples $\{y_{ij}\}$ were used in the Gibbs sampler and after 3000 “burn-in” period of iterations, 1000 iterations were kept for analysis. By using the Gibbs sampling output, we calculated pseudo posterior estimators $\hat{\theta}_i^p$ and $\hat{V}^p(\theta_i)$ based on (5.24) and (5.25) and model-based posterior estimators $\hat{\theta}_i^m$ and $\hat{V}^m(\theta_i)$ based on (5.27) and (5.28), respectively. By substituting the census estimates $\bar{\sigma}_e^2$ and $\bar{\sigma}_v^2$ for σ_e^2 and σ_v^2 in (5.8), we computed the “census” posterior mean of θ_i as $\bar{\theta}_i^c = E(\theta_i | \bar{Y}_w, \bar{\sigma}_e^2, \bar{\sigma}_v^2)$. The “census” posterior variance of θ_i was calculated using (5.9) as $\bar{V}^c(\theta_i) = V(\theta_i | \bar{Y}_w, \bar{\sigma}_e^2, \bar{\sigma}_v^2)$. Note that $\bar{\theta}_i^c$ and $\bar{V}^c(\theta_i)$ were calculated only for simulation purpose, since in practice usually it is impossible to obtain the census estimates of σ_e^2 and σ_v^2 . This whole process was repeated $R = 100$ times, and for each run r ($r = 1, \dots, R$), $\bar{y}_{iw}(r)$, $\hat{\theta}_i^p(r)$, $\hat{\theta}_i^m(r)$, and $\bar{\theta}_i^c(r)$ were computed, as well as the associated posterior variances $\hat{V}^p(\theta_i)(r)$, $\hat{V}^m(\theta_i)(r)$, and $\bar{V}^c(\theta_i)(r)$. This simulation

approach is design-based in the sense that it refers to the design and the PPS samples were drawn from a fixed population.

Our purpose of this simulation study is to test the performance of the proposed pseudo posterior estimators. For this purpose, we calculated the absolute relative bias (ARB) of $\hat{\theta}_i^p$ with respect to the fixed small area mean θ_i as

$$ARB(\hat{\theta}_i^p) = \left| \frac{E^*(\hat{\theta}_i^p) - \theta_i}{\theta_i} \right| = |E^*(\hat{\theta}_i^p)/\theta_i - 1|,$$

where E^* denotes expectation with respect to the design, that is, the average value over $R = 100$ runs, $E^*(\hat{\theta}_i^p) = \frac{1}{R} \sum_{r=1}^R \hat{\theta}_i^p(r)$. Similarly the absolute relative bias of $\hat{\theta}_i^m$ was computed as $ARB(\hat{\theta}_i^m) = |E^*(\hat{\theta}_i^m)/\theta_i - 1|$.

The empirical coefficient of variation (CV) of $\hat{\theta}_i^p$ and $\hat{\theta}_i^m$ were calculated as

$$CV(\hat{\theta}_i^p) = E^*(\sqrt{\hat{V}^p(\theta_i)}/\hat{\theta}_i^p),$$

and

$$CV(\hat{\theta}_i^m) = E^*(\sqrt{\hat{V}^m(\theta_i)}/\hat{\theta}_i^m).$$

The relative efficiency (RE) of $\hat{\theta}_i^p$ and $\hat{\theta}_i^m$ over the direct survey estimator $\bar{y}_{iw}$ were computed as

$$RE(\hat{\theta}_i^p) = MSE^*(\bar{y}_{iw})/MSE^*(\hat{\theta}_i^p),$$

and

$$RE(\hat{\theta}_i^m) = MSE^*(\bar{y}_{iw})/MSE^*(\hat{\theta}_i^m),$$

where MSE^* denotes the MSE over $R = 100$ runs. For example, $MSE^*(\hat{\theta}_i^p) = \sum_{r=1}^R (\hat{\theta}_i^p(r) - \theta_i)^2 / R$, where θ_i is the fixed small area mean of the i th area.

The relative difference (RD) of $\hat{\theta}_i^p$ and $\tilde{\theta}_i^c$ was computed as

$$\begin{aligned} RD[\hat{\theta}_i^p, \tilde{\theta}_i^c] &= E^*(|\hat{\theta}_i^p - \tilde{\theta}_i^c|/\tilde{\theta}_i^c) \\ &= \frac{1}{R} \sum_{r=1}^R |\hat{\theta}_i^p(r) - \tilde{\theta}_i^c(r)|/\tilde{\theta}_i^c(r). \end{aligned}$$

Similarly the relative difference of $\hat{V}^p(\theta_i)$ and $\tilde{V}^c(\theta_i)$ was computed as

$$RD[\hat{V}^p(\theta_i), \tilde{V}^c(\theta_i)] = E^*(|\hat{V}^p(\theta_i) - \tilde{V}^c(\theta_i)|/\tilde{V}^c(\theta_i)).$$

And finally we also calculated the ratio $\hat{V}^n(\theta_i)/\hat{V}^p(\theta_i)$ over R runs as

$$\hat{V}^n(\theta_i)/\hat{V}^p(\theta_i) = \frac{1}{R} \sum_{r=1}^R \hat{V}^n(\theta_i)(r)/\hat{V}^p(\theta_i)(r).$$

Tables 5.1, 5.2 and 5.3 present the average values of ARB, CV and RE for $\hat{\theta}_i^p$ and $\hat{\theta}_i^m$ over the small areas $i = 1, \dots, 30$ for the three populations. First of all, the ARB is very small for both $\hat{\theta}_i^p$ and $\hat{\theta}_i^m$, even when the sample size n is small. For example, when $n = 5$, $\hat{\theta}_i^p$ has about 7% and $\hat{\theta}_i^m$ has about 6% ARB. As sample size n increases, the ARB decreases for both $\hat{\theta}_i^p$ and $\hat{\theta}_i^m$. The pseudo posterior mean $\hat{\theta}_i^p$ has slightly larger bias than the posterior mean $\hat{\theta}_i^m$ due to the bias induced by the direct survey estimate $\bar{y}_{iw}$ in $\hat{\theta}_i^p$. For the CV, $\hat{\theta}_i^p$ has about 2% larger CV than $\hat{\theta}_i^m$ due to the extra uncertainty associated with the $\bar{y}_{iw}$. So in terms of ARB and CV, $\hat{\theta}_i^p$ performs very much like the $\hat{\theta}_i^m$, except for the extra bias and uncertainty induced by $\bar{y}_{iw}$ for $\hat{\theta}_i^p$. Since the sample obeys the assumed model, $\hat{\theta}_i^m$ performs well.

For the RE, both $\hat{\theta}_i^p$ and $\hat{\theta}_i^m$ have large RE gain over $\bar{y}_{iw}$. As sample size n decreases, the RE gain increases. For example, in Table 5.1, $RE(\hat{\theta}_i^p) = 1.2$ if $n = 40$; $RE(\hat{\theta}_i^p) = 1.3$ if $n = 20$; $RE(\hat{\theta}_i^p) = 2.0$ if $n = 5$. So the model-assisted estimators indeed borrowing strength from related areas and the gains in efficiency over $\bar{y}_{iw}$

are more significant when the sample size is small. This is because when n is small, the survey estimator $\bar{y}_{iw}$ is not reliable, thus the improvement of the model-assisted estimator is more evident. When the sample size is the same, $\hat{\theta}_i^m$ has larger RE gain than $\hat{\theta}_i^p$ as expected. Note that as sample size increases, the pseudo posterior mean $\hat{\theta}_i^p$ moves toward $\bar{y}_{iw}$ (design-consistent), whereas $\hat{\theta}_i^m$ moves toward the sample mean $\bar{y}_i = \frac{1}{n_i} \sum_j y_{ij}$. Given that the model is true, we know that $\bar{y}_i$ is the BLUE of θ_i and is always better than $\bar{y}_{iw}$. Thus, as it is shown in the results, as n increases, $RE(\hat{\theta}_i^p)$ decreases and moves toward 1, whereas $RE(\hat{\theta}_i^m)$ is always larger than $RE(\hat{\theta}_i^p)$. Overall, the pseudo posterior mean $\hat{\theta}_i^p$ has less RE gain than $\hat{\theta}_i^m$. This is the trade-off for bringing the survey weights into the model and making $\hat{\theta}_i^p$ design-consistent.

The RE gain also depends on the ratio σ_v^2/σ_e^2 . When the sample size is fixed, as σ_v^2/σ_e^2 increases, the weight r_{iw} in $\hat{\theta}_i^p$ will increase and more weight will be put on $\bar{y}_{iw}$ so that the RE gain of $\hat{\theta}_i^p$ over $\bar{y}_{iw}$ decreases. For example, when $n = 20$, under Pop 3 ($\sigma_v^2/\sigma_e^2 = 0.16$), $RE(\hat{\theta}_i^p) = 2.0$; under Pop 2 ($\sigma_v^2/\sigma_e^2 = 0.28$), $RE(\hat{\theta}_i^p) = 1.6$; under Pop 1 ($\sigma_v^2/\sigma_e^2 = 0.44$), $RE(\hat{\theta}_i^p) = 1.3$.

Table 5.4 shows the comparison of $\hat{\theta}_i^p$ and $\tilde{\theta}_i^c$. The RD of $\hat{\theta}_i^p$ and $\tilde{\theta}_i^c$ is very small and negligible, even when the sample size is as small as $n = 5$. However, for the RD of $\hat{V}^p(\theta_i)$ and $\tilde{V}^c(\theta_i)$, when n is small, the RD can be very large. As n increases, the RD decreases. Thus it is clear that $\hat{V}^p(\theta_i)$ takes into account of the extra uncertainty associated with the estimation of σ_v^2 and σ_e^2 .

Table 5.5 shows the results on the ratio $\hat{V}^n(\theta_i)/\hat{V}^p(\theta_i)$. The underestimation of $\hat{V}^p(\theta_i)$ induced by using $\hat{V}^n(\theta_i)$ is very small: when $n \geq 20$, $\hat{V}^n(\theta_i)/\hat{V}^p(\theta_i) \geq 97.6\%$

for all three populations. Even when $n = 5$, the values of $\hat{V}^n(\theta_i)/\hat{V}^p(\theta_i)$ under the three populations are 95.3%, 93.3% and 91.5%, respectively. This is because (1) the number of small areas $m = 30$ is relatively large and (2) the unit-level model is used to estimate σ_e^2 and σ_v^2 , and the uncertainty of estimating σ_e^2 and σ_v^2 is relatively small with respect to the posterior variance based on the aggregated area-level model.

In summary, the pseudo posterior mean $\hat{\theta}_i^p$ is a good point estimator for θ_i in terms of ARB and CV. As the small area sample size increases, $\hat{\theta}_i^p$ is design-consistent. The pseudo posterior variance $\hat{V}^p(\theta_i)$ is an appropriate measure of uncertainty associated with $\hat{\theta}_i^p$.

5.5 Extension to Multi-level Models

We have proposed a Bayesian approach to obtain a pseudo hierarchical Bayes estimator of a small area mean under simple random effects models. This pseudo Bayesian estimator depends on the survey weights and is design-consistent. We also obtained the pseudo posterior variance of the small area mean. In this section, we extend the pseudo Bayes approach to the multi-level models considered in Chapter 4 to make use of survey weights and borrow strength across areas.

Let us consider the unequal error variance multi-level model:

$$y_{ij} = x_{ij}^T \beta_i + e_{ij}, \quad e_{ij} \sim N(0, \sigma_i^2), \quad (j = 1, \dots, n_i; i = 1, \dots, m); \quad (5.29)$$

where β_i is a random regression coefficient vector modelled as

$$\beta_i = Z_i \gamma + v_i, \quad v_i \sim N_p(0, \Phi), \quad (i = 1, \dots, m), \quad (5.30)$$

where γ is unknown regression parameters, v_i is a random effect vector, area-specific

error variance σ_i^2 is unknown for all i , and variance covariance matrix Φ is also unknown.

Let $\tilde{w}_{ij}$ be the basic design weight attached to y_{ij} and $w_{ij} = \tilde{w}_{ij} / \sum_j \tilde{w}_{ij}$ be the normalized survey weight ($\sum_j w_{ij} = 1$). Then combining the survey weights and the unit-level model (5.29), we can get the following area-level model

$$\bar{y}_{iw} = \sum_{j=1}^{n_i} w_{ij} y_{ij} = \sum_{j=1}^{n_i} w_{ij} (x_{ij}^T \beta_i + e_{ij}) = \bar{x}_{iw}^T \beta_i + \bar{e}_{iw} \quad (5.31)$$

where $\bar{x}_{iw} = \sum_{j=1}^{n_i} w_{ij} x_{ij}$ and $\bar{e}_{iw} = \sum_{j=1}^{n_i} w_{ij} e_{ij}$ with $E(\bar{e}_{iw}) = 0$ and $V(\bar{e}_{iw}) = \sigma_i^2 \sum_{j=1}^{n_i} w_{ij}^2 = \sigma_i^2 d_i$.

Let $\tau_i = \sigma_i^{-2}$ and $\Omega = \Phi^{-1}$, the conditional distribution of β_i given the survey estimate $\bar{y}_{iw}$ and parameters τ_i , γ , and Ω is obtained as

$$[\beta_i | \bar{y}_{iw}, \tau_i, \gamma, \Omega] \sim N(M^{-1}(\Omega Z_i \gamma + \tau_i d_i^{-1} \bar{x}_{iw} \bar{y}_{iw}), M^{-1}) \quad (5.32)$$

where $M = \tau_i d_i^{-1} \bar{x}_{iw} \bar{x}_{iw}^T + \Omega$. Then the posterior mean and posterior variance of β_i are given as follows:

$$\begin{aligned} \tilde{\beta}_i &= E(\beta_i | \bar{y}_{iw}, \tau_i, \gamma, \Omega) \\ &= M^{-1}(\Omega Z_i \gamma + \tau_i d_i^{-1} \bar{x}_{iw} \bar{y}_{iw}) \\ &= (\tau_i d_i^{-1} \bar{x}_{iw} \bar{x}_{iw}^T + \Omega)^{-1}(\Omega Z_i \gamma + \tau_i d_i^{-1} \bar{x}_{iw} \bar{y}_{iw}) \end{aligned} \quad (5.33)$$

and

$$\tilde{V}(\beta_i) = V(\beta_i | \bar{y}_{iw}, \tau_i, \gamma, \Omega) = (\tau_i d_i^{-1} \bar{x}_{iw} \bar{x}_{iw}^T + \Omega)^{-1}. \quad (5.34)$$

Since τ_i , γ and Ω are all unknown, we implement the Gibbs sampler based on the unit multi-level model (5.29) and draw samples for τ_i , γ and Ω from the following full conditional distributions:

(i) For $i = 1, \dots, m$, $[\beta_i|Y, \gamma, \Omega, \tau] \stackrel{ind}{\sim}$

$$N_p((\tau_i X_i^T X_i + \Omega)^{-1}(\tau_i X_i^T Y_i + \Omega Z_i \gamma), (\tau_i X_i^T X_i + \Omega)^{-1})$$

(ii) $[\gamma|Y, \beta, \Omega, \tau] \sim$

$$N_q((\sum_{i=1}^m Z_i^T \Omega Z_i + D^{-1})^{-1}(\sum_{i=1}^m Z_i^T \Omega \beta_i), (\sum_{i=1}^m Z_i^T \Omega Z_i + D^{-1})^{-1})$$

(iii) $[\Omega|Y, \beta, \gamma, \tau] \sim W_p(\alpha + m, R + \frac{1}{2} \sum_{i=1}^m (\beta_i - Z_i \gamma)(\beta_i - Z_i \gamma)^T)$

(iv) For $i = 1, \dots, m$,

$$[\tau_i|Y, \beta, \gamma, \Omega] \sim G(a_i + \frac{1}{2}n_i, b_i + \frac{1}{2}(Y_i - X_i \beta_i)^T(Y_i - X_i \beta_i))$$

where $Y = (\{y_{ij}\}, j = 1, \dots, n_i, i = 1, \dots, m)$, D, α, R, a_i and b_i are known prior parameters (see Chapter 4). Suppose a sample of size G is obtained as $\{\tau_i^{(k)}, i = 1, \dots, m; \gamma^{(k)}; \Omega^{(k)}\}, k = 1, \dots, G$, then the pseudo posterior mean of β_i is obtained as:

$$\begin{aligned} \hat{\beta}_i^p &= \frac{1}{G} \sum_{k=1}^G E(\beta_i | \bar{y}_{iw}, \tau_i^{(k)}, \gamma^{(k)}, \Omega^{(k)}) \\ &= \frac{1}{G} \sum_{k=1}^G (\tau_i^{(k)} d_i^{-1} \bar{x}_{iw} \bar{x}_{iw}^T + \Omega^{(k)})^{-1} (\Omega^{(k)} Z_i \gamma^{(k)} + \tau_i^{(k)} d_i^{-1} \bar{x}_{iw} \bar{y}_{iw}). \end{aligned} \quad (5.35)$$

We refer to $\hat{\beta}_i^p$ as the pseudo Rao-Blackwellized estimator of β_i .

Similarly, we obtained the pseudo Rao-Blackwellized estimator of the posterior variance of β_i as:

$$\begin{aligned} \hat{V}^p(\beta_i) &= \frac{1}{G} \sum_{k=1}^G V(\beta_i | \bar{y}_{iw}, \tau_i^{(k)}, \gamma^{(k)}, \Omega^{(k)}) + \frac{1}{G} \sum_{k=1}^G [E(\beta_i | \bar{y}_{iw}, \tau_i^{(k)}, \gamma^{(k)}, \Omega^{(k)})]^2 \\ &\quad - [\frac{1}{G} \sum_{k=1}^G E(\beta_i | \bar{y}_{iw}, \tau_i^{(k)}, \gamma^{(k)}, \Omega^{(k)})]^2 \\ &= \frac{1}{G} \sum_{k=1}^G (\tau_i^{(k)} d_i^{-1} \bar{x}_{iw} \bar{x}_{iw}^T + \Omega^{(k)})^{-1} \end{aligned}$$

$$\begin{aligned}
& + \frac{1}{G} \sum_{k=1}^G (\tau_i^{(k)} d_i^{-1} \bar{x}_{iw} \bar{x}_{iw}^T + \Omega^{(k)})^{-1} (\Omega^{(k)} Z_i \gamma^{(k)} + \tau_i^{(k)} d_i^{-1} \bar{x}_{iw} \bar{y}_{iw}) \\
& \quad \times (\Omega^{(k)} Z_i \gamma^{(k)} + \tau_i^{(k)} d_i^{-1} \bar{x}_{iw} \bar{y}_{iw})^T (\tau_i^{(k)} d_i^{-1} \bar{x}_{iw} \bar{x}_{iw}^T + \Omega^{(k)})^{-1} \\
& - \frac{1}{G^2} \left[\sum_{k=1}^G (\tau_i^{(k)} d_i^{-1} \bar{x}_{iw} \bar{x}_{iw}^T + \Omega^{(k)})^{-1} (\Omega^{(k)} Z_i \gamma^{(k)} + \tau_i^{(k)} d_i^{-1} \bar{x}_{iw} \bar{y}_{iw}) \right] \\
& \quad \times \left[\sum_{k=1}^G (\tau_i^{(k)} d_i^{-1} \bar{x}_{iw} \bar{x}_{iw}^T + \Omega^{(k)})^{-1} (\Omega^{(k)} Z_i \gamma^{(k)} + \tau_i^{(k)} d_i^{-1} \bar{x}_{iw} \bar{y}_{iw}) \right]^T. \quad (5.36)
\end{aligned}$$

Finally the estimator of the i th small area mean μ_i is obtained as:

$$\begin{aligned}
\hat{\mu}_i^p &= \bar{X}_i^T \hat{\beta}_i^p \\
&= \frac{1}{G} \sum_{k=1}^G \bar{X}_i^T (\tau_i^{(k)} d_i^{-1} \bar{x}_{iw} \bar{x}_{iw}^T + \Omega^{(k)})^{-1} (\Omega^{(k)} Z_i \gamma^{(k)} + \tau_i^{(k)} d_i^{-1} \bar{x}_{iw} \bar{y}_{iw}), \quad (5.37)
\end{aligned}$$

and its measure of uncertainty is given by:

$$\hat{V}^p(\mu_i) = \bar{X}_i^T \cdot \hat{V}^p(\beta_i) \cdot \bar{X}_i \quad (5.38)$$

where $\hat{V}^p(\beta_i)$ is given by (5.36), and $\bar{X}_i$ is the population mean of the auxiliary variable for the i th small area. Note that $\hat{\mu}_i^p$ and $\hat{V}^p(\mu_i)$ depend on the survey weights and borrow strength across areas. Also $\hat{\mu}_i^p$ is design-consistent when the small area sample size increases.

Population 1: ($\sigma_e^2 = 225, \sigma_v^2 = 100$)

Measure	$n = 5$	$n = 20$	$n = 40$
$ARB(\hat{\theta}_i^p) \%$	6.7	3.5	2.3
$ARB(\hat{\theta}_i^m) \%$	5.9	3.1	2.1
$CV(\hat{\theta}_i^p) \%$	12.5	8.5	6.7
$CV(\hat{\theta}_i^m) \%$	11.0	6.3	4.6
$RE(\hat{\theta}_i^p)$	2.0	1.3	1.2
$RE(\hat{\theta}_i^m)$	2.5	2.2	2.0

Table 5.1: Comparison of $\hat{\theta}_i^p$ and $\hat{\theta}_i^m$ Under Population 1Population 2: ($\sigma_e^2 = 225, \sigma_v^2 = 64$)

Measure	$n = 5$	$n = 20$	$n = 40$
$ARB(\hat{\theta}_i^p) \%$	7.0	3.6	2.4
$ARB(\hat{\theta}_i^m) \%$	6.1	3.1	2.0
$CV(\hat{\theta}_i^p) \%$	11.7	8.2	6.6
$CV(\hat{\theta}_i^m) \%$	10.6	6.2	4.6
$RE(\hat{\theta}_i^p)$	2.9	1.6	1.4
$RE(\hat{\theta}_i^m)$	3.2	2.6	2.3

Table 5.2: Comparison of $\hat{\theta}_i^p$ and $\hat{\theta}_i^m$ Under Population 2

Population 3: ($\sigma_e^2 = 225, \sigma_y^2 = 36$)

Measure	$n = 5$	$n = 20$	$n = 40$
$ARB(\hat{\theta}_i^p) \%$	7.2	4.1	2.6
$ARB(\hat{\theta}_i^m) \%$	6.3	3.2	1.9
$CV(\hat{\theta}_i^p) \%$	9.3	7.6	6.4
$CV(\hat{\theta}_i^m) \%$	8.5	5.9	4.4
$RE(\hat{\theta}_i^p)$	3.9	2.0	1.6
$RE(\hat{\theta}_i^m)$	4.5	2.9	2.5

Table 5.3: Comparison of $\hat{\theta}_i^p$ and $\hat{\theta}_i^m$ Under Population 3

	$RD(\hat{\theta}_i^p, \bar{\theta}_i^c) \%$		
Population	$n = 5$	$n = 20$	$n = 40$
Population 1	1.1	0.4	0.1
Population 2	1.5	0.5	0.2
Population 3	1.6	0.5	0.2
	$RD(\hat{V}^p(\theta_i), \bar{V}^c(\theta_i)) \%$		
Population	$n = 5$	$n = 20$	$n = 40$
Population 1	10.2	4.8	3.0
Population 2	16.3	6.1	3.6
Population 3	22.1	7.6	4.9

Table 5.4: Relative difference of $\hat{\theta}_i^p$ and $\bar{\theta}_i^c$

Population	$n = 5$	$n = 20$	$n = 40$
Population 1	0.953	0.987	0.992
Population 2	0.933	0.984	0.990
Population 3	0.915	0.976	0.986

Table 5.5: Results of $\hat{V}^n(\theta_i)/\hat{V}^p(\theta_i)$

Chapter 6

Bayesian Analysis of Nonlinear Mixed Models in Small Area Estimation

6.1 Introduction

Linear mixed models including simple random effects models, nested error regression models and multi-level models are usually used in small area estimation when response variables are continuous. We have studied the HB approach for such models in Chapters 4 and 5. For categorical response variables, like binary variables, generalized linear models with random effects including mixed logistic and probit regression models have been used in small area estimation under a Bayesian framework (Ghosh, et al., 1998; Farrell, et al., 1997). Recently Jiang and Lahiri (1998) developed the Prasad-Rao type frequentist alternative to the Bayesian approach for small area estimation with binary data. They obtained the EB estimator and an approximation to its MSE correct to terms of order m^{-1} for the case of a binary variable and logistic regression model with a random effect.

In this Chapter, we consider two types of nonlinear models: (i) a unit-level nonlinear model and (ii) an area-level nonlinear model. For both models, the linking model for the population mean is a log linear model. For the unit-level model, following the approach used in Chapter 5 for linear models, we use survey weights to form an aggregated area-level model. Then following Jiang and Lahiri (1998), the predictor of random effect is first obtained assuming regression parameters and variance components are known. Afterwards, pseudo posterior estimators of small area means are obtained by using the method in Chapter 5. Metropolis-Hastings algorithm within Gibbs sampler is used to generate samples because we cannot get closed form conditional distributions for some parameters due to nonlinearity of the model.

For the area-level model, two approaches are considered. One is a completely hierarchical Bayes method using Markov chain Monte Carlo sampling method including Gibbs sampling and Metropolis-Hastings algorithm. In the second approach, log transformation is applied to the direct estimates to match the linking model, leading to a Fay-Herriot type linear model (Fay and Herriot, 1979). Estimators of the transformed small area means are first obtained under the Fay-Herriot model, and then transformed back to get estimates of small area means. But, bias in estimation and underestimation of the posterior variance under the true model might be introduced by this transformation approach. A simulation study is conducted to compare the results of these two approaches.

6.2 A Unit-level Log Linear Model

Consider the following unit-level small area model:

$$y_{ij} = \theta_i + e_{ij}, j = 1, \dots, N_i; i = 1, \dots, m, \quad (6.1)$$

and the linking model for the small area mean θ_i :

$$\log(\theta_i) = x_i^T \beta + v_i, i = 1, \dots, m, \quad (6.2)$$

where y_{ij} are the unit population values and $e_{ij} \sim N(0, \sigma_e^2)$, N_i is the population size in the i -th area, and x_i is an area-level auxiliary variable ($i = 1, \dots, m$). The v_i 's are independent normal random variables with $E(v_i) = 0$ and $V(v_i) = \sigma_v^2$. Further, v_i and e_{ij} are assumed to be uncorrelated for all i and j . For simplicity, we take θ_i as the small area mean $\bar{Y}_i = \sum_j y_{ij}/N_i$ because $\theta_i \approx \bar{Y}_i$ if N_i is large.

6.2.1 Predictor of random effect

Assuming models (6.1) and (6.2) also hold for the sample $\{y_{ij}, j = 1, \dots, n_i, i = 1, \dots, m\}$, we get a nonlinear random effect model as

$$y_{ij} = \theta_i + e_{ij} = e^{x_i^T \beta + v_i} + e_{ij}, j = 1, \dots, n_i, i = 1, \dots, m. \quad (6.3)$$

Let $\tilde{w}_{ij}$ denote the basic design weight attached to y_{ij} . As in Chapter 5, let $w_{ij} = \tilde{w}_{ij} / \sum_j \tilde{w}_{ij}$ be the normalized weights and

$$\bar{y}_{iw} = \frac{\sum_{j=1}^{n_i} \tilde{w}_{ij} y_{ij}}{\sum_{j=1}^{n_i} \tilde{w}_{ij}} = \sum_{j=1}^{n_i} w_{ij} y_{ij} \quad (6.4)$$

be the direct estimator of θ_i . By combining the direct estimator (6.4) with the sampling model (6.3), we get the following aggregated area-level model:

$$\bar{y}_{iw} = \sum_{j=1}^{n_i} w_{ij} (\theta_i + e_{ij})$$

$$\begin{aligned}
&= \theta_i + \bar{e}_{iw} \\
&= e^{x_i^T \beta} e^{v_i} + \bar{e}_{iw},
\end{aligned} \tag{6.5}$$

where $\bar{e}_{iw} = \sum_{j=1}^{n_i} w_{ij} e_{ij}$ with $E(\bar{e}_{iw}) = 0$ and $V(\bar{e}_{iw}) = \sigma_e^2 \sum_{j=1}^{n_i} w_{ij}^2 \equiv \sigma_e^2 d_i$.

Assuming the parameters β , σ_e^2 and σ_v^2 are known, the best predictor of e^{v_i} given the survey estimate $\bar{y}_{iw}$ is given by

$$\begin{aligned}
E(e^{v_i} | \bar{y}_{iw}, \beta, \sigma_e^2, \sigma_v^2) &= \frac{\int e^{v_i} f(\bar{y}_{iw} | v_i) f(v_i) dv_i}{\int f(\bar{y}_{iw} | v_i) f(v_i) dv_i} \\
&= \frac{E \exp(\eta + \phi_i(\beta, \sigma_e^2, \eta))}{E \exp(\phi_i(\beta, \sigma_e^2, \eta))}.
\end{aligned} \tag{6.6}$$

where

$$\phi_i(\beta, \sigma_e^2, \eta) = -\frac{(\bar{y}_{iw} - e^{x_i^T \beta + \eta})^2}{2\sigma_e^2 d_i} \tag{6.7}$$

and $\eta \sim N(0, \sigma_v^2)$. Numerical evaluation of (6.6) involves only one dimensional integration with respect to a normal random variable. However, since equation (6.6) does not have a closed form, Monte Carlo integration is needed in numerical computation.

By using $V(x) = E(x^2) - (E(x))^2$, we can also find the conditional variance of e^{v_i} given $\bar{y}_{iw}$ and β , σ_e^2 , σ_v^2 as

$$\begin{aligned}
V(e^{v_i} | \bar{y}_{iw}, \beta, \sigma_e^2, \sigma_v^2) &= E(e^{2v_i} | \bar{y}_{iw}, \beta, \sigma_e^2, \sigma_v^2) - [E(e^{v_i} | \bar{y}_{iw}, \beta, \sigma_e^2, \sigma_v^2)]^2 \\
&= \frac{E \exp(2\eta + \phi_i(\beta, \sigma_e^2, \eta))}{E \exp(\phi_i(\beta, \sigma_e^2, \eta))} - \left[\frac{E \exp(\eta + \phi_i(\beta, \sigma_e^2, \eta))}{E \exp(\phi_i(\beta, \sigma_e^2, \eta))} \right]^2.
\end{aligned} \tag{6.8}$$

6.2.2 Pseudo Bayes estimator

We are interested in obtaining the posterior estimator of the small area mean θ_i .

Under the area-level model (6.5), it follows from the predictor (6.6) that

$$E(\theta_i | \bar{y}_{iw}, \beta, \sigma_e^2, \sigma_v^2) = E(e^{x_i^T \beta + v_i} | \bar{y}_{iw}, \beta, \sigma_e^2, \sigma_v^2)$$

$$\begin{aligned}
&= e^{x_i^T \beta} \cdot E(e^{v_i} | \bar{y}_{iw}, \beta, \sigma_e^2, \sigma_v^2) \\
&= e^{x_i^T \beta} \frac{E \exp(\eta + \phi_i(\beta, \sigma_e^2, \eta))}{E \exp(\phi_i(\beta, \sigma_e^2, \eta))} \\
&\equiv \bar{\theta}_i(\beta, \sigma_e^2, \sigma_v^2).
\end{aligned} \tag{6.9}$$

The posterior variance of θ_i given $\bar{y}_{iw}$ and $\beta, \sigma_e^2, \sigma_v^2$ is obtained as

$$\begin{aligned}
V(\theta_i | \bar{y}_{iw}, \beta, \sigma_e^2, \sigma_v^2) &= V(e^{x_i^T \beta + v_i} | \bar{y}_{iw}, \beta, \sigma_e^2, \sigma_v^2) \\
&= e^{2x_i^T \beta} \cdot V(e^{v_i} | \bar{y}_{iw}, \beta, \sigma_e^2, \sigma_v^2) \\
&\equiv \bar{V}_{\theta_i}(\beta, \sigma_e^2, \sigma_v^2),
\end{aligned} \tag{6.10}$$

where $V(e^{v_i} | \bar{y}_{iw}, \beta, \sigma_e^2, \sigma_v^2)$ is given by (6.8).

The estimators $\bar{\theta}_i$ and $\bar{V}_{\theta_i}$ depend on the unknown parameters β, σ_e^2 , and σ_v^2 . Let $\pi(\beta, \sigma_e^2, \sigma_v^2 | Y)$ denote the joint posterior distribution of β, σ_e^2 , and σ_v^2 from the unit-level data $Y = \{y_{ij}, j = 1, \dots, n_i, i = 1, \dots, m\}$. Following Chapter 5, we obtained the pseudo posterior mean and pseudo posterior variance of θ_i as follows:

$$\begin{aligned}
\theta_i^p &= E_{\beta, \sigma_e^2, \sigma_v^2} \bar{\theta}_i(\beta, \sigma_e^2, \sigma_v^2) \\
&= \int \bar{\theta}_i(\beta, \sigma_e^2, \sigma_v^2) \pi(\beta, \sigma_e^2, \sigma_v^2 | Y) d\beta d\sigma_e^2 d\sigma_v^2 \\
&= \int E(\theta_i | \bar{y}_{iw}, \beta, \sigma_e^2, \sigma_v^2) \pi(\beta, \sigma_e^2, \sigma_v^2 | Y) d\beta d\sigma_e^2 d\sigma_v^2 \\
&\equiv \theta_i^p(\bar{y}_{iw}, Y)
\end{aligned} \tag{6.11}$$

and

$$\begin{aligned}
V_{\theta_i}^p &= E_{\beta, \sigma_e^2, \sigma_v^2} \bar{V}_{\theta_i} + V_{\beta, \sigma_e^2, \sigma_v^2} \bar{\theta}_i \\
&= \int \bar{V}_{\theta_i}(\beta, \sigma_e^2, \sigma_v^2) \pi(\beta, \sigma_e^2, \sigma_v^2 | Y) d\beta d\sigma_e^2 d\sigma_v^2 \\
&+ \int \bar{\theta}_i^2 \pi(\beta, \sigma_e^2, \sigma_v^2 | Y) d\beta d\sigma_e^2 d\sigma_v^2 - [\int \bar{\theta}_i \pi(\beta, \sigma_e^2, \sigma_v^2 | Y) d\beta d\sigma_e^2 d\sigma_v^2]^2 \\
&\equiv V_{\theta_i}^p(\bar{y}_{iw}, Y)
\end{aligned} \tag{6.12}$$

Since $\tilde{\theta}_i$ and $\tilde{V}_{\theta_i}$ do not have closed forms, evaluation of θ_i^p in (6.11) and $V_{\theta_i}^p$ in (6.12) by direct integration is not feasible. Thus the Gibbs sampling method is used to obtain Monte Carlo estimates of θ_i^p and $V_{\theta_i}^p$.

6.2.3 Monte Carlo evaluation

By assuming a flat prior on β and inverse gamma priors on σ_e^2 and σ_v^2 , i.e., $\sigma_e^2 \sim IG(a_e, b_e)$ and $\sigma_v^2 \sim IG(a_v, b_v)$, where a_e, b_e, a_v, b_v are known and usually chosen to be very small, we obtain the following full conditional distributions for the Gibbs sampler to simulate samples from $\pi(\beta, \sigma_e^2, \sigma_v^2 | Y)$:

$$\theta_i | Y, \beta, \sigma_e^2, \sigma_v^2 \propto \frac{1}{\theta_i} \exp \left\{ -\frac{\sum_{j=1}^{n_i} (y_{ij} - \theta_i)^2}{2\sigma_e^2} - \frac{(\log \theta_i - x_i^T \beta)^2}{2\sigma_v^2} \right\}; \quad (6.13)$$

$$\beta | Y, \theta, \sigma_e^2, \sigma_v^2 \sim N \left(\left(\sum_{i=1}^m x_i x_i^T \right)^{-1} \left(\sum_{i=1}^m x_i \log \theta_i \right), \sigma_v^2 \left(\sum_{i=1}^m x_i x_i^T \right)^{-1} \right); \quad (6.14)$$

$$\sigma_e^2 | Y, \theta, \beta, \sigma_v^2 \sim IG \left(a_e + \frac{1}{2} \sum_{i=1}^m n_i, b_e + \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \theta_i)^2 \right); \quad (6.15)$$

$$\sigma_v^2 | Y, \theta, \beta, \sigma_e^2 \sim IG \left(a_v + \frac{1}{2} m, b_v + \frac{1}{2} \sum_{i=1}^m (\log \theta_i - x_i^T \beta)^2 \right). \quad (6.16)$$

To implement the Gibbs sampler, we need to update θ , β , σ_e^2 and σ_v^2 iteratively by drawing samples from the conditional distributions (6.13) to ((6.16). Sampling from the conditional distributions of β , σ_e^2 and σ_v^2 is straightforward. However, the conditional distribution $[\theta_i | Y, \beta, \sigma_e^2, \sigma_v^2]$ does not have a closed form. Thus the Metropolis-Hastings algorithm, discussed in Chapter 2, is used to generate samples from $[\theta_i | Y, \beta, \sigma_e^2, \sigma_v^2]$. For this we note that the density function of $[\theta_i | Y, \beta, \sigma_e^2, \sigma_v^2]$ can be written as

$$\pi(\theta_i | Y, \beta, \sigma_e^2, \sigma_v^2) \propto \psi(\theta_i, \sigma_e^2) \cdot f(\theta_i | \beta, \sigma_v^2), \quad (6.17)$$

where $f(\theta_i|\beta, \sigma_v^2)$ is a log normal density function given as

$$f(\theta_i|\beta, \sigma_v^2) = \frac{1}{\sqrt{2\pi}\sigma_v\theta_i} \exp\left\{-\frac{1}{2\sigma_v^2}(\log\theta_i - x_i^T\beta)^2\right\}, \quad (6.18)$$

and $\psi(\theta_i, \sigma_e^2)$ is given as

$$\psi(\theta_i, \sigma_e^2) = \exp\left\{-\frac{\sum_{j=1}^{n_i}(y_{ij} - \theta_i)^2}{2\sigma_e^2}\right\}. \quad (6.19)$$

To implement the Metropolis-Hastings algorithm within the Gibbs sampler, we use $f(\theta_i|\beta, \sigma_v^2)$ as the candidate generating density function in the Metropolis-Hastings sampling step. Then the Metropolis probability of move at the k th iteration is given by

$$\alpha(\theta_i^{(k)}, \theta_i^{(k+1)}) = \min\left\{\frac{\psi(\theta_i^{(k+1)}, \sigma_e^{2(k)})}{\psi(\theta_i^{(k)}, \sigma_e^{2(k)})}, 1\right\}. \quad (6.20)$$

The sampling procedure is summarized as follows: Suppose the Markov chain is currently at $\{\theta^{(k)}, \beta^{(k)}, \sigma_e^{2(k)}, \sigma_v^{2(k)}\}$, then to move to the next step, we proceed as follows:

- For $i = 1, \dots, m$, to update θ_i , we draw a candidate $\theta_i^{(k+1)}$ from the log normal density $f(\theta_i|\beta^{(k)}, \sigma_v^{2(k)})$. With probability $\alpha(\theta_i^{(k)}, \theta_i^{(k+1)})$ given by (6.20), this candidate is accepted; otherwise set $\theta_i^{(k+1)} = \theta_i^{(k)}$;
- draw $\beta^{(k+1)}$ from $[\beta|Y, \theta^{(k+1)}, \sigma_e^{2(k)}, \sigma_v^{2(k)}]$;
- draw $\sigma_e^{2(k+1)}$ from $[\sigma_e^2|Y, \theta^{(k+1)}, \beta^{(k+1)}, \sigma_v^{2(k)}]$;
- draw $\sigma_v^{2(k+1)}$ from $[\sigma_v^2|Y, \theta^{(k+1)}, \beta^{(k+1)}, \sigma_e^{2(k)}]$.

Suppose a sample $\{\beta^{(k)}, \sigma_e^{2(k)}, \sigma_v^{2(k)}; k = 1, \dots, G\}$ of size G is available from the above Gibbs sampler. Then it follows from (6.11) and (6.12) that the estimated

pseudo posterior mean and posterior variance of the small area mean θ_i are:

$$\begin{aligned}\hat{\theta}_i^p &= \frac{1}{G} \sum_{k=1}^G E(\theta_i | \bar{y}_{iw}, \beta^{(k)}, \sigma_e^{2(k)}, \sigma_v^{2(k)}) \\ &= \frac{1}{G} \sum_{k=1}^G \tilde{\theta}_i(\beta^{(k)}, \sigma_e^{2(k)}, \sigma_v^{2(k)})\end{aligned}\quad (6.21)$$

and

$$\begin{aligned}\hat{V}^p(\theta_i) &= \frac{1}{G} \sum_{k=1}^G \tilde{V}_{\theta_i}(\beta^{(k)}, \sigma_e^{2(k)}, \sigma_v^{2(k)}) + \frac{1}{G} \sum_{k=1}^G [\tilde{\theta}_i(\beta^{(k)}, \sigma_e^{2(k)}, \sigma_v^{2(k)})]^2 \\ &\quad - \left[\frac{1}{G} \sum_{k=1}^G \tilde{\theta}_i(\beta^{(k)}, \sigma_e^{2(k)}, \sigma_v^{2(k)}) \right]^2,\end{aligned}\quad (6.22)$$

where $\tilde{\theta}_i$ and $\tilde{V}_{\theta_i}$ are given by (6.9) and (6.10), respectively. To evaluate each $\tilde{\theta}_i(\beta^{(k)}, \sigma_e^{2(k)}, \sigma_v^{2(k)})$ and $\tilde{V}_{\theta_i}(\beta^{(k)}, \sigma_e^{2(k)}, \sigma_v^{2(k)})$, Monte Carlo integration is needed since we do not have closed forms for $\tilde{\theta}_i$ and $\tilde{V}_{\theta_i}$. This is extra computational cost compared to the case of linear models considered in Chapter 5.

6.2.4 Remarks

The unit-level models given by (6.1) and (6.2) can be extended to the following model where unit-level auxiliary variables can be used:

$$y_{ij} = \theta_{ij} + e_{ij}, j = 1, \dots, N_i; i = 1, \dots, m, \quad (6.23)$$

and

$$\log(\theta_{ij}) = x_{ij}^T \beta + v_i, i = 1, \dots, m, \quad (6.24)$$

where x_{ij} is the unit-level auxiliary variable. Then jointly we have

$$y_{ij} = e^{x_{ij}^T \beta + v_i} + e_{ij},$$

and the i th small area mean is given by

$$\bar{Y}_i = \left(\sum_{j=1}^{N_i} e^{x_{ij}^T \beta} \right) e^{v_i}, \quad (6.25)$$

assuming the population size N_i is large.

Given the sample $\{y_{ij}\}$ and the normalized weights $\{w_{ij}\}$, following (6.5), we will have the following aggregated model

$$\bar{y}_{iw} = \left(\sum_{j=1}^{n_i} w_{ij} e^{x_{ij}^T \beta} \right) e^{v_i} + \bar{e}_{iw}. \quad (6.26)$$

Assuming the parameters β , σ_e^2 and σ_v^2 are known, and conditional on the survey estimate $\bar{y}_{iw}$, the posterior estimator of the small area mean $\bar{Y}_i$ can be obtained as

$$\begin{aligned} E(\bar{Y}_i | \bar{y}_{iw}, \beta, \sigma_e^2, \sigma_v^2) &= E\left(\left(\sum_{j=1}^{N_i} e^{x_{ij}^T \beta}\right) e^{v_i} | \bar{y}_{iw}, \beta, \sigma_e^2, \sigma_v^2\right) \\ &= \left(\sum_{j=1}^{N_i} e^{x_{ij}^T \beta}\right) \cdot E(e^{v_i} | \bar{y}_{iw}, \beta, \sigma_e^2, \sigma_v^2) \\ &= \left(\sum_{j=1}^{N_i} e^{x_{ij}^T \beta}\right) \frac{E \exp(\eta + \phi_i(\beta, \sigma_e^2, \eta))}{E \exp(\phi_i(\beta, \sigma_e^2, \eta))} \end{aligned} \quad (6.27)$$

where

$$\phi_i(\beta, \sigma_e^2, \eta) = -\frac{(\bar{y}_{iw} - (\sum_{j=1}^{n_i} w_{ij} e^{x_{ij}^T \beta}) e^\eta)^2}{2\sigma_e^2 d_i} \quad (6.28)$$

and $\eta \sim N(0, \sigma_v^2)$. The posterior variance of $\bar{Y}_i$ given $\bar{y}_{iw}$ and β , σ_e^2 , σ_v^2 can be obtained as

$$\begin{aligned} V(\bar{Y}_i | \bar{y}_{iw}, \beta, \sigma_e^2, \sigma_v^2) &= V\left(\left(\sum_{j=1}^{N_i} e^{x_{ij}^T \beta}\right) e^{v_i} | \bar{y}_{iw}, \beta, \sigma_e^2, \sigma_v^2\right) \\ &= \left(\sum_{j=1}^{N_i} e^{x_{ij}^T \beta}\right)^2 \cdot V(e^{v_i} | \bar{y}_{iw}, \beta, \sigma_e^2, \sigma_v^2). \end{aligned} \quad (6.29)$$

Following Section 6.2.2, we can similarly obtain the pseudo posterior mean and pseudo posterior variance of $\bar{Y}_i$. For Monte Carlo evaluation and implementation of the Gibbs sampler, the following full conditional distributions are needed:

$$\beta|Y, v_i, \sigma_e^2, \sigma_v^2 \propto \exp\left\{-\frac{1}{2\sigma_e^2} \sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - e^{x_{ij}^T \beta + v_i})^2\right\}; \quad (6.30)$$

$$v_i|Y, \theta, \sigma_e^2, \sigma_v^2 \propto \exp\left\{-\frac{1}{2\sigma_e^2} \sum_{j=1}^{n_i} (y_{ij} - e^{x_{ij}^T \beta + v_i})^2 - \frac{1}{2\sigma_v^2} v_i^2\right\}; \quad (6.31)$$

$$\sigma_e^2|Y, \beta, v_i, \sigma_v^2 \sim IG\left(a_e + \frac{1}{2} \sum_{i=1}^m n_i, b_e + \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - e^{x_{ij}^T \beta + v_i})^2\right); \quad (6.32)$$

$$\sigma_v^2|Y, \beta, v_i, \sigma_e^2 \sim IG\left(a_v + \frac{1}{2} m, b_v + \frac{1}{2} \sum_{i=1}^m v_i^2\right). \quad (6.33)$$

Metropolis algorithm within the Gibbs sampler is needed for sampling from both β and v_i , unlike the case of area-level auxiliary variables. In particular, sampling from β needs more work. Usually a multivariate normal distribution with mean equal to the mode of $f(\beta|Y, v_i, \sigma_e^2, \sigma_v^2)$ and variance covariance matrix equal to the expected curvature at the mode is used as the candidate generating density function in the Metropolis algorithm (see Chapter 2, Section 2.3; Zeger and Karim (1991). Czado (1994)).

6.3 An Area-level Model

In this section, we consider a nonlinear area-level model for small area estimation. Suppose there are m small areas. Let y_i denote the direct survey estimator of the i th small area mean θ_i , $i = 1, \dots, m$. Consider the following area-level nonlinear mixed model given as:

the sampling model:

$$y_i|\theta_i \sim N(\theta_i, \theta_i^2 D_i), \quad i = 1, \dots, m, \quad (6.34)$$

and the linking model:

$$\log \theta_i | \beta, \sigma^2 \sim N(x_i^T \beta, \sigma^2), \quad (i = 1, \dots, m), \quad (6.35)$$

where D_i is known, β and σ^2 are unknown. We cannot combine (6.34) and (6.35) to obtain a linear model.

It is important to note that in the sampling model (6.34), the sampling error variance of y_i depends on the unknown small area mean θ_i . The coefficient of variation(CV) of y_i is $\sqrt{D_i}$ and does not depend θ_i , thus the standard deviation increases in direct proportion to the expected value θ_i . This kind of model was considered in Fay and Herriot (1979) for the application of James-Stein estimator.

The model (6.35) for the small area mean θ_i is a log linear model and is equivalent to

$$\theta_i | \beta, \sigma^2 \sim \log N(x_i^T \beta, \sigma^2), \quad (6.36)$$

i.e., θ_i has a log normal distribution with density function given by

$$f(\theta_i | \beta, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma\theta_i} \exp\left\{-\frac{1}{2\sigma^2}(\log\theta_i - x_i^T \beta)^2\right\}. \quad (6.37)$$

We are interested in estimating the small area mean θ_i . Two approaches will be considered in the following subsections.

6.3.1 A transformation approach

In the model (6.34), the sampling variance of y_i is proportional to θ_i^2 and is unknown since θ_i is unknown. Logarithmic transformation is usually used to stabi-

lize the variance of y_i and then the sampling variances could be considered known (Fay and Herriot, 1979). Then the combined model is a linear Fay-Herriot model. The resulting estimates are then transformed back to the original scale. Recently, county-level (area-level) models with logarithmic transformation have been used to produce model-based county-level estimates of school-age children in poverty in the United States (National Research Council, 1997, 1998). Using these estimates, the US Department of Education allocates \$7 billion of federal funds to counties, and states distribute these funds among school districts in each county. Variables are transformed using logarithms for two reasons. First, it is more plausible that the model is homoscedastic on the log scale (corresponding to a constant coefficient of variation or equal model variances of share in poverty) than on the original scale over the extremely wide range of county sizes. Second, the transformed variables have a much more symmetric distribution, and the scatterplots of various covariates with the dependent variable are more linear.

For the transformation approach, let $z_i = \log(y_i)$ and $\mu_i = \log(\theta_i)$. Then approximately we get $E(z_i) \approx \mu_i$ and $Var(z_i) \approx D_i$. Ignoring the approximation leads to the well known Fay-Herriot model:

$$z_i | \mu_i \sim N(\mu_i, D_i), \quad i = 1, \dots, m; \quad (6.38)$$

and

$$\mu_i | \beta, \sigma^2 \sim N(x_i^T \beta, \sigma^2), \quad i = 1, \dots, m. \quad (6.39)$$

This model is based on the transformed data z_i and parameter μ_i . We are interested in $E(\theta_i | Y)$ and $Var(\theta_i | Y)$. Since $\theta_i = e^{\mu_i}$, the first step is to obtain the

estimator of μ_i . The Gibbs sampling method is used to generate samples from the posterior distribution of μ_i . Suppose we have a sample $\{\mu_i^{(k)}, k = 1, \dots, G\}$ of size G generated from the Gibbs sampler, accordingly we have $\{\theta_i^{(k)}, k = 1, \dots, G\}$ with $\theta_i^{(k)} = e^{\mu_i^{(k)}}$. Then $E(\theta_i|Y)$ and $Var(\theta_i|Y)$ are estimated by the sample mean and sample variance of $\{\theta_i^{(k)}, k = 1, \dots, G\}$, respectively.

Assuming a flat prior on β and $\sigma^2 \sim IG(a, b)$, where a and b are known, we obtained the following full conditional distributions for the Gibbs sampler based on the transformed model:

(1) For $i = 1, \dots, m$,

$$[\mu_i|Y, \beta, \sigma^2] \sim N((1 - r_i)y_i + r_i x_i^T \beta, D_i(1 - r_i)),$$

$$\text{where } r_i = \frac{D_i}{\sigma^2 + D_i};$$

(2) $[\beta|Y, \mu, \sigma^2] \sim N((\sum_{i=1}^m x_i x_i^T)^{-1} (\sum_{i=1}^m x_i \mu_i), \sigma^2 (\sum_{i=1}^m x_i x_i^T)^{-1});$

(3) $[\sigma^2|Y, \mu, \beta] \sim IG(a + \frac{1}{2}m, b + \frac{1}{2} \sum_{i=1}^m (\mu_i - x_i^T \beta)^2).$

Since all the conditional distributions have closed-form, implementing of the Gibbs sampler is straightforward. Thus by using logarithmic transformation, the estimation procedure is simplified. However, bias could be introduced via transformation because the approximation may not be accurate for small sample sizes. In the next section, we consider a fully Bayesian approach based on the original nonlinear model to evaluate this bias.

6.3.2 A complete sampling approach

Consider the original area-level models (6.34) and (6.35). We are interested in finding the posterior mean and posterior variance of θ_i 's given the data $Y = (y_1, \dots, y_m)^T$.

From models (6.34) and (6.35), the joint density function of Y and parameters θ, β, σ^2 is given by

$$f(Y, \theta, \beta, \sigma^2) \propto \prod_{i=1}^m \frac{1}{\theta_i} \exp\left\{-\frac{(y_i - \theta_i)^2}{2\theta_i^2 D_i}\right\} \cdot \prod_{i=1}^m \frac{1}{\sigma \theta_i} \exp\left\{-\frac{(\log \theta_i - x_i^T \beta)^2}{2\sigma^2}\right\} \cdot \pi(\beta, \sigma^2), \quad (6.40)$$

where $\pi(\beta, \sigma^2)$ is a prior distribution of β and σ^2 . Then the joint posterior distribution of θ_i 's is given by

$$f(\theta|Y) \propto \int f(Y, \theta, \beta, \sigma^2) d\beta d\sigma^2.$$

Direct evaluation of the joint posterior distribution involves high-dimensional integration, and is not computationally feasible. Thus we use the Gibbs sampler to generate samples from the posterior distribution of θ_i 's.

The joint distribution (6.40) can be employed to determine the following full conditional distributions required by the Gibbs sampler:

(1) For $i = 1, \dots, m$,

$$[\theta_i|Y, \beta, \sigma^2] \propto \frac{1}{\theta_i^2} \exp\left\{-\frac{(y_i - \theta_i)^2}{2\theta_i^2 D_i} - \frac{(\log \theta_i - x_i^T \beta)^2}{2\sigma^2}\right\};$$

(2) $[\beta|Y, \theta, \sigma^2] \sim N((\sum_{i=1}^m x_i x_i^T)^{-1} (\sum_{i=1}^m x_i \log \theta_i), \sigma^2 (\sum_{i=1}^m x_i x_i^T)^{-1});$

(3) $[\sigma^2|Y, \theta, \beta] \sim IG(a + \frac{1}{2}m, b + \frac{1}{2} \sum_{i=1}^m (\log \theta_i - x_i^T \beta)^2).$

The conditional distribution $[\theta_i|Y, \beta, \sigma^2]$ does not have a closed form, Metropolis-Hastings algorithm is used to generate samples from $[\theta_i|Y, \beta, \sigma^2]$. We note that the conditional density function of $[\theta_i|Y, \beta, \sigma^2]$ can be written as

$$\pi(\theta_i|Y, \beta, \sigma^2) \propto h(\theta_i) \cdot f(\theta_i|\beta, \sigma^2) \quad (6.41)$$

where $f(\theta_i|\beta, \sigma^2)$ is the log normal density function given by (6.37) and $h(\theta_i)$ is a function given by

$$h(\theta_i) = \frac{1}{\theta_i} \exp\left\{-\frac{(y_i - \theta_i)^2}{2\theta_i^2 D_i}\right\}. \quad (6.42)$$

To implement the Metropolis-Hastings sampling step within the Gibbs sampler, we use $f(\theta_i|\beta, \sigma^2)$ as the candidate generating density function. Then at the k th iteration, the Metropolis probability of move is given by

$$\alpha(\theta_i^{(k)}, \theta_i^{(k+1)}) = \min\left\{\frac{h(\theta_i^{(k+1)})}{h(\theta_i^{(k)})}, 1\right\}. \quad (6.43)$$

Suppose the Markov chain is currently at $\{\theta^{(k)}, \beta^{(k)}, \sigma^{2(k)}\}$, the Metropolis-Hastings algorithm within the Gibbs sampler can be summarized as follows:

- For $i = 1, \dots, m$, to update θ_i , we draw a candidate $\theta_i^{(k+1)}$ from the log normal density $f(\theta_i|\beta^{(k)}, \sigma^{2(k)})$. With probability $\alpha(\theta_i^{(k)}, \theta_i^{(k+1)})$ given by (6.43), this candidate is accepted; otherwise set $\theta_i^{(k+1)} = \theta_i^{(k)}$.
- To update β , draw $\beta^{(k+1)}$ from $[\beta|Y, \theta^{(k+1)}, \sigma^{2(k)}]$.
- To update σ^2 , draw $\sigma^{2(k+1)}$ from $[\sigma^2|Y, \theta^{(k+1)}, \beta^{(k+1)}]$.

The estimation of $E(\theta_i|Y)$ and $V(\theta_i|Y)$ is based on the marginal sample $\{\theta_i^{(k)}, k = 1, \dots, G\}$ from the Gibbs sampler. We will compare this direct Gibbs sampling approach with the logarithmic transformation approach through a numerical study in the next section.

6.4 A Numerical Study

In this section, we compare the results of the two approaches introduced in Section 6.3, i.e., the transformation approach and the complete Gibbs sampling approach, for

the area-level nonlinear model given by (6.34) and (6.35). We are mainly interested in the relative differences of the posterior estimates based on the two approaches. For this, we set the number of small areas $m = 12$, and consider the following random effect area-level model:

$$y_i \sim N(\theta_i, \theta_i^2 D_i), \quad i = 1, \dots, 12; \quad (6.44)$$

$$\log(\theta_i) = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + v_i, \quad v_i \sim N(0, \sigma_v^2); \quad (6.45)$$

where the design matrix is chosen as follows:

i	1	2	3	4	5	6	7	8	9	10	11	12
x_{i1}	1	1	1	1	1	1	0	0	0	0	0	0
x_{i2}	-6	-5	-4	-3	-2	-1	1	2	3	4	5	6

The fixed effects coefficients are set at $\beta_0 = -1.5$, $\beta_1 = 1.0$, $\beta_2 = 0.5$. Variance component $\sigma_v^2 = 1.5$, and $D_i = 1/n$, where n in practice is usually considered to be the sample size for the i th small area. In our study, n is chosen to be 5, 10, 20, 50, and 100, respectively.

For each n , we generated 10 data sets according to models (6.44) and (6.45). Let $\hat{\theta}_i^T$ denote the estimator of the posterior mean of θ_i based on the transformation approach, and $\hat{\theta}_i^G$ denote the corresponding estimator based on the complete Gibbs sampling approach. For both approaches, the Gibbs sampler was run for 10,000 iterations with the first 5,000 iterations as “burn-in” period, and the last 5,000 iterations were kept for analysis. For each data set, we calculated $\hat{\theta}_i^T$ and $\hat{\theta}_i^G$, and the corresponding posterior variances $\hat{V}_{\theta_i}^T$ and $\hat{V}_{\theta_i}^G$, where “T” denotes the transformation approach and “G” denotes the complete Gibbs sampling approach. The relative

difference (RD) of the two estimators $\hat{\theta}_i^T$ and $\hat{\theta}_i^G$ is computed as

$$RD_i = \frac{|\hat{\theta}_i^G - \hat{\theta}_i^T|}{\hat{\theta}_i^G}. \quad (6.46)$$

and the ratio of the two posterior variances (RV) as

$$RV_i = \frac{\hat{V}_{\theta_i}^G}{\hat{V}_{\theta_i}^T}. \quad (6.47)$$

Table 6.1 presents the relative differences of the two posterior mean estimators based on the 10 data sets for the 12 synthetic small areas. Summary measures include the minimum (Min), maximum (Max) and mean (Mean) over the 12 small areas. For each area, as n increases, or as D_i decreases, the relative difference becomes smaller. In particular, when $n = 100$, or $D_i = 0.01$, the RD_i is between 0.45% to 0.99%, the average over the areas is only 0.68%, even less than 1%. When $n = 20$ or $D_i = 0.05$, the RD_i is between 2.81% to 4.41%, the average is 3.61%, which is less than 5%. Whereas when $n = 5$ or $D_i = 0.2$, the relative difference is quite large, ranging from 9.59% to 27.68%, and the average is about 18.4%.

Table 6.2 presents the ratio of the two posterior variances. Table 6.2 shows that when $n = 100$ or $D_i = 0.01$, the two posterior variances, $\hat{V}_{\theta_i}^T$ and $\hat{V}_{\theta_i}^G$, are very close to each other, with the ratio $\hat{V}_{\theta_i}^G/\hat{V}_{\theta_i}^T$ ranging from 0.9925 to 1.1350, and the average is about 1.045. When $n = 20$, the ratio $\hat{V}_{\theta_i}^G/\hat{V}_{\theta_i}^T$ ranges from 1.4125 to 1.8942, and the average is 1.5931. When $n = 5$, the ratio could be very large, and the average is about 6.2696.

If $D_i = 0.01$ or the CV of y_i is 10%, then the two approaches are approximately equivalent to each other with approximately equal posterior mean estimators and posterior variance estimators, i.e., $\hat{\theta}_i^T \approx \hat{\theta}_i^G$ and $\hat{V}_{\theta_i}^T \approx \hat{V}_{\theta_i}^G$. If $D_i = 0.02$ or the CV

of y_i is about 14.14%, $\hat{V}_{\theta_i}^T$ is about 87.27% of $\hat{V}_{\theta_i}^G$. If the CV of y_i is about 22.36%, then $\hat{V}_{\theta_i}^T$ is about 62.77% of $\hat{V}_{\theta_i}^G$. Thus in practice, care should be taken when using logarithmic transformation on the original data. When the CV of y_i is less than 10%, the logarithmic transformation approach works quite well. If the CV is about 15%, the logarithmic transformation could lead to about 15% underestimation of the posterior variance compared to the complete Gibbs sampling approach, even though the posterior mean estimation is about the same. However, if the CV of y_i is large, say, over 20%, the transformation approach could lead to serious underestimation of the posterior variance and relatively large bias of the posterior mean estimator.

Area	$n = 5$	$n = 10$	$n = 20$	$n = 50$	$n = 100$
1	0.2440	0.1131	0.0348	0.0125	0.0060
2	0.1292	0.0763	0.0334	0.0125	0.0099
3	0.1828	0.0667	0.0253	0.0091	0.0081
4	0.1976	0.0847	0.0366	0.0086	0.0049
5	0.0959	0.0532	0.0284	0.0131	0.0069
6	0.1883	0.0786	0.0320	0.0088	0.0066
7	0.2030	0.1118	0.0265	0.0132	0.0056
8	0.1344	0.0674	0.0287	0.0110	0.0075
9	0.1687	0.0940	0.0261	0.0086	0.0077
10	0.2768	0.1425	0.0441	0.0165	0.0090
11	0.2031	0.0860	0.0281	0.0101	0.0045
12	0.1840	0.0865	0.0347	0.0104	0.0051
Min	0.0959	0.0532	0.0281	0.0091	0.0045
Max	0.2768	0.1425	0.0441	0.0165	0.0099
Mean	0.1840	0.0884	0.0316	0.0112	0.0068

Table 6.1: Relative Differences of Posterior Means

Area	$n = 5$	$n = 10$	$n = 20$	$n = 50$	$n = 100$
1	7.5041	3.7927	1.7838	1.1652	1.0544
2	4.5603	3.2831	1.5016	1.1080	0.9959
3	2.9238	2.1268	1.4034	1.1415	1.0687
4	4.0929	2.3879	1.7403	1.1895	1.1093
5	2.1300	1.7318	1.4125	1.1779	0.9925
6	10.0006	3.1884	1.6649	1.1360	1.0349
7	6.4311	2.0022	1.4791	1.1254	1.0238
8	3.6106	2.7955	1.5145	1.1354	1.0487
9	5.7843	3.1406	1.5554	1.1177	1.1350
10	8.7815	6.5512	1.8942	1.1848	1.0565
11	9.4120	3.0060	1.5678	1.1678	1.0087
12	10.0047	3.4418	1.5992	1.1017	0.9993
Min	2.1300	1.7318	1.4125	1.1017	0.9925
Max	10.0047	6.5512	1.8942	1.1895	1.1350
Mean	6.2696	3.1207	1.5931	1.1459	1.0450

Table 6.2: Ratio of Posterior Variances

Chapter 7

Hierarchical Bayes Analysis of Small Area Binary Data

7.1 Introduction

In this Chapter, we consider binary data analysis for small area estimation. Recently, some work has appeared on Bayesian analysis of binary survey data. Farrell et al. (1997) and MacGibbison and Tomberlin (1989) have obtained small area estimates of proportions via the empirical Bayes method. Farrell (1995) used the hierarchical Bayes approach to estimate small area proportions, and Malec et al. (1997) obtained predictive inference on a linear combination of binary variables using a logistic regression (logit) model with random regression coefficients and a HB approach. Stroud (1994) provided a comprehensive Bayesian analysis of binary survey data under various sampling designs. Ghosh et al. (1998) provided a general MCMC sampling HB approach for small area estimation using generalized linear models. In light of the previous work, we consider only two types of problems in this Chapter: (i) Estimate small area characteristics such as small area means or totals by combining binary survey data and small area models; (ii) Smooth small

area direct survey estimates through hierarchical models. Thus this Chapter is divided into two main sections. Section 7.2 provides general techniques for small area estimation using binary data and predictive inference (Malec et al., 1997). We proposed a two-step sampling scheme and a random effect probit model, instead of a logit model, for which latent variables can be introduced for the Gibbs sampling method. Section 7.3 provides a compound beta-binomial model in the case of multi-level nesting and absence of auxiliary variables with application to smoothing direct estimates of response rates in the Canadian Homeowner Repair and Renovation Survey (HRRS).

7.2 Small Area Estimation

Suppose there are m small areas. Let y_{ij} denote a binary random variable for the j th unit of the i th small area, $i = 1, \dots, m; j = 1, \dots, N_i$, where N_i is the population size of units for the i th area. We are interested in making inference about finite population quantites, specifically a linear function of the y_{ij} 's, i.e.,

$$\theta = \sum_{i=1}^m \sum_{j=1}^{N_i} a_{ij} y_{ij}, \quad (7.1)$$

where a_{ij} 's are specified constants that can be chosen to make θ as the population mean, or the population total or a subpopulation parameter.

7.2.1 Hierarchical probit model

For the binary variable y_{ij} , we consider the following hierarchical Bayes probit random effect model:

(I) Conditional on p_{ij} , y_{ij} are assumed to be independent Bernoulli random variables, i.e.,

$$y_{ij}|p_{ij} \sim \text{Bernoulli}(p_{ij}); \quad (7.2)$$

(II) Population model for p_{ij} :

$$\Phi^{-1}(p_{ij}) = x_{ij}^T \beta + v_i, \quad v_i \stackrel{iid}{\sim} N(0, \sigma_v^2); \quad (7.3)$$

(III) Prior distributions: $\beta \sim N_q(0, D)$, and $\sigma_v^2 \sim IG(\frac{1}{2}a, \frac{1}{2}b)$, $a > 0, b > 0$.

Here Φ is the standard normal distribution function, β is the vector of fixed effect probit regression parameters, x_{ij} is the vector of unit-level auxiliary variables, and $IG(\alpha, \beta)$ denotes an inverse gamma distribution with density function given by $f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-(\alpha+1)} e^{-\frac{\beta}{x}}$, $\alpha > 0, \beta > 0, x > 0$. and D, a, b are known constants.

Let Y_s denote the vector of sampled units. We are particularly interested in estimating the posterior mean and posterior variance of θ given the sample observations Y_s , i.e., $E(\theta|Y_s)$ and $Var(\theta|Y_s)$. From (7.1) and (7.2), we have

$$\begin{aligned} E(\theta|Y_s) &= E\left(\sum_{i=1}^m \sum_{j=1}^{N_i} a_{ij} y_{ij} | Y_s\right) \\ &= \sum_{(i,j) \in s} a_{ij} y_{ij} + \sum_{(i,j) \notin s} a_{ij} E(p_{ij}|Y_s) \end{aligned} \quad (7.4)$$

and

$$\begin{aligned} Var(\theta|Y_s) &= \sum_{(i,j) \notin s} a_{ij}^2 E(p_{ij}(1-p_{ij})|Y_s) + Var\left(\sum_{(i,j) \notin s} a_{ij} p_{ij} | Y_s\right) \\ &= \sum_{(i,j) \notin s} a_{ij}^2 E(p_{ij}(1-p_{ij})|Y_s) + E\left(\sum_{(i,j) \notin s} a_{ij} p_{ij} | Y_s\right)^2 \\ &\quad - \left(\sum_{(i,j) \notin s} a_{ij} E(p_{ij}|Y_s)\right)^2, \end{aligned} \quad (7.5)$$

where $(i, j) \in s$ means that unit j of the i th small area is in the sample and $p_{ij} = \Phi(x_{ij}^T \beta + v_i)$.

7.2.2 Inferential methodology

Since $E(\theta|Y_s)$ and $V(\theta|Y_s)$ depend on β and $v_i, i \notin s$ through a nonlinear function p_{ij} , and the posterior distribution

$$f(\beta, v_i, (i \notin s), \sigma_v^2 | Y_s)$$

does not have a simple form, Monte Carlo numerical evaluation is needed. Suppose we generate G sets of samples $\{\beta^{(k)}, v_i^{(k)}, i \notin s, \sigma_v^{2(k)}\}, k = 1, \dots, G$ from the posterior distribution, then we can evaluate $p_{ij}^{(k)}$ as

$$p_{ij}^{(k)} = \Phi(x_{ij}^T \beta^{(k)} + v_i^{(k)}), (i, j) \notin s. \quad (7.6)$$

Then estimators of $E(\theta|Y_s)$ and $Var(\theta|Y_s)$ can be constructed respectively as

$$\hat{\theta} = \sum_{(i,j) \in s} a_{ij} y_{ij} + \frac{1}{G} \sum_{k=1}^G \left[\sum_{(i,j) \notin s} a_{ij} p_{ij}^{(k)} \right] \quad (7.7)$$

and

$$\begin{aligned} \hat{V}_{\theta} &= \frac{1}{G} \sum_{k=1}^G \left[\sum_{(i,j) \notin s} a_{ij}^2 p_{ij}^{(k)} (1 - p_{ij}^{(k)}) + \left(\sum_{(i,j) \notin s} a_{ij} p_{ij}^{(k)} \right)^2 \right] \\ &\quad - \left(\frac{1}{G} \sum_{k=1}^G \left(\sum_{(i,j) \notin s} a_{ij} p_{ij}^{(k)} \right) \right)^2. \end{aligned} \quad (7.8)$$

To draw samples from the joint posterior distribution of β, v_i and σ_v^2 , we introduce latent variables z_{ij} as follows:

$$y_{ij} = \begin{cases} 1 & \text{if } z_{ij} > 0 \\ 0 & \text{if } z_{ij} \leq 0 \end{cases}$$

where

$$z_{ij} | \beta, v_i \sim N(x_{ij}^T \beta + v_i, 1). \quad (7.9)$$

Let $v = (v_1, \dots, v_m)$ and $Z = (z_{ij}, i = 1, \dots, m; j = 1, \dots, N_i)$, then it is easy to prove the following lemma:

Lemma 7.1: The joint posterior distribution of β, v, σ_v^2, Z given Y_s can be represented as follows:

$$\begin{aligned} f(\beta, v, \sigma_v^2, Z | Y_s) &= f_1(z_{ij}, v_i, (i, j) \in s; \beta, \sigma_v^2 | Y_s) f_2(v_i, i \notin s | \sigma_v^2) \\ &\quad \cdot f_3(z_{ij}, i = 1, \dots, m, j \notin s | v, \beta, \sigma_v^2). \end{aligned} \quad (7.10)$$

According to Lemma 7.1 and following Malec et al. (1997), we generate samples in two steps:

- (1) apply Gibbs sampling procedure to f_1 in (7.10) to get $\{\beta^{(k)}, \sigma_v^{2(k)}, v_i^{(k)}, i \in s\}$;
- (2) generate $\{v_i^{(k)}, i \notin s\}$ directly from f_2 in (7.10) using $\sigma_v^{2(k)}$ from stage (1).

Then we have the required $\{\beta^{(k)}, v_i^{(k)}, i = 1, \dots, m\}$, and the inference is based on $\beta^{(k)}$ and $v_i^{(k)}$.

To implement the Gibbs sampler in step (1), we need the following full conditional distributions derived from $f_1(z_{ij}, v_i, (i, j) \in s, \beta, \sigma_v^2 | Y_s)$:

Algorithm 1:

- (i) $[\beta | z_{ij}, v_i, (i, j) \in s, \sigma_v^2, Y_s] \sim N((X^T X + D^{-1})^{-1}(X^T Z_s - \sum_{i \in s} v_i \sum_{j \in s} x_{ij}), (X^T X + D^{-1})^{-1}),$
where Z_s is a vector containing $z_{ij}, (i, j) \in s$, and $X^T = \text{col}_{(i,j) \in s}(x_{ij})$.
- (ii) $[\sigma_v^2 | z_{ij}, v_i, (i, j) \in s, \beta, Y_s] \sim IG(\frac{1}{2}(a + \sum_{i \in s} 1), \frac{1}{2}(b + \sum_{i \in s} v_i^2)).$

(iii) For $i \in s$, $[v_i | z_{ij}, (i, j) \in s, \beta, \sigma_v^2, Y_s] \sim$

$$N((n_i + \frac{1}{\sigma_v^2})^{-1} \sum_{j \in s} (z_{ij} - x_{ij}^T \beta), (n_i + \frac{1}{\sigma_v^2})^{-1}),$$

where n_i is the sample size of i th small area with $i \in s$.

(iv) $[z_{ij} | v_i, (i, j) \in s, \beta, \sigma_v^2, Y_s] \sim \begin{cases} N(x_{ij}^T \beta + v_i, 1), z_{ij} > 0 & \text{if } y_{ij} = 1 \\ N(x_{ij}^T \beta + v_i, 1), z_{ij} \leq 0 & \text{if } y_{ij} = 0 \end{cases}$

Since all the full conditional distributions have closed forms, unlike in the logit model case, implementation of the Gibbs sampler is straightforward. Conditionally z_{ij} has a truncated normal distribution. There are several algorithms available to simulate truncated normal variables; for instance, see Robert (1995). Chip and Carlin (1997) introduced the idea of blocking parameters in MCMC sampling to improve the performance of the Gibbs sampler. The algorithm may be refined by marginalizing the distribution of latent variable z_{ij} over the random effect v_i . Then

$$z_{ij} \sim N(x_{ij}^T \beta, 1 + \sigma_v^2).$$

The resulting algorithm is similar to Algorithm 1 except that β is sampled from $\beta | z_{ij}, \sigma_v^2$ and z_{ij} from $\sim N(x_{ij}^T \beta, 1 + \sigma_v^2)$ truncated by y_{ij} . We summarize this algorithm as follows:

Algorithm 2:

(i) $[\beta | z_{ij}, \sigma_v^2, Y_s] \sim N(((1 + \sigma_v^2)^{-1} X^T X + D^{-1})^{-1} X^T Z_s, ((1 + \sigma_v^2)^{-1} X^T X + D^{-1})^{-1}).$

(ii) $[\sigma_v^2 | z_{ij}, v_i, (i, j) \in s, \beta, Y_s] \sim IG(\frac{1}{2}(a + \sum_{i \in s} 1), \frac{1}{2}(b + \sum_{i \in s} v_i^2)).$

(iii) For $i \in s$, $[v_i | z_{ij}, (i, j) \in s, \beta, \sigma_v^2, Y_s] \sim$

$$N((n_i + \frac{1}{\sigma_v^2})^{-1} \sum_{j \in s} (z_{ij} - x_{ij}^T \beta), (n_i + \frac{1}{\sigma_v^2})^{-1}).$$

$$(iv) [z_{ij}|v_i, (i, j) \in s, \beta, \sigma_v^2, Y_s] \sim \begin{cases} N(x_{ij}^T \beta, 1 + \sigma_v^2), z_{ij} > 0 & \text{if } y_{ij} = 1 \\ N(x_{ij}^T \beta, 1 + \sigma_v^2), z_{ij} \leq 0 & \text{if } y_{ij} = 0 \end{cases}$$

Then for the binary case, Algorithm 2 may be the best overall, in terms of both ease of use and quality of the output (Chip and Carlin, 1997).

We have provided only a sketch of the theory using hierarchical probit models. We plan to implement this method on real survey data at Statistics Canada.

7.3 Smoothing Small Area Estimates

7.3.1 Background

Introduced in 1987, Statistics Canada's Homeowner Repair and Renovation Survey (HRRS) is a supplementary survey to the Canadian Labour Force Survey (LFS). The LFS and HRRS follow the same sample design which is a multistage self-weighting design. The LFS is conducted every month whereas HRRS is an annual survey conducted in March. The LFS provides estimates of labour force characteristics whereas HRRS measures levels of expenditure by Canadian homeowners for additions, renovations, repairs and maintenance, and installation and replacement of household equipment. The LFS sample follows a rotating sample design and is composed of six rotation groups of approximately equal size. The HRRS uses two thirds of the LFS sample and therefore uses four rotation groups. The household expenditure information for the HRRS is collected from the LFS respondents. Thus the HRRS nonrespondent stands for LFS respondent and HRRS nonrespondent, and the HRRS response rate is about 88% of the LFS respondents at the national level. Currently the HRRS nonresponse is adjusted for by the same method as is used for LFS nonre-

sponse: the sample is divided into response homogeneity groups (RHG) within each province, in most cases, an RHG is the intersection of an economic region, a frame type, and a rotation, and weights within each RHG are multiplied by the inverse of its weighted response rate. This method produces somewhat unstable HRRS nonresponse adjustment factors, due to the small size of many of the RHG's. This makes it necessary to use appropriate models and auxiliary information to borrow strength from related regions to find more accurate estimate for the response rate or to smooth the empirical estimate of the response rate. For this purpose, we propose a hierarchical Bayes model to derive a model-based estimator of the response rate.

Hierarchical Bayes method has been proposed in the literature to smooth the direct survey estimates. Datta et al. (1998) obtained hierarchical Bayes estimates of unemployment rates for the U.S. states using a cross sectional and time series model. Maiti (1998) proposed a hierarchical Bayes model involving Poisson distribution and a log-linear model to smooth the relative risks of a disease for small areas. He and Sun (1998) used a beta-binomial model to obtain posterior success rates for small areas. Rosner (1989) presented multivariate methods for clustered binary data with more than one level of nesting. In particular, a compound beta-binomial model that generalizes the beta-binomial distribution to more than one level of nesting was presented in the absence of auxiliary covariates. In this section, a compound beta-binomial model will be used to model the direct survey estimate of the response rate of HRRS. A hierarchical Bayes approach is presented in smoothing the direct survey estimates. The Gibbs sampling method is used to obtain the posterior estimate of the response rate and posterior variance is used as the measure of uncertainty

associated with the estimate.

7.3.2 Hierarchical beta-binomial model

Let p denote the national level response rate, let p_i denote the provincial level response rate for the i th province, and let p_{ij} denote the response rate of the j th response homogeneity group (RHG) in the i th province. Furthermore, let n_{ij} be the sample size of the ij th RHG, and y_{ij} the number of respondents among the n_{ij} sample units. We consider the following hierarchical Bayes compound beta-binomial model:

Sampling model:

$$y_{ij}|p_{ij} \sim \text{Binomial}(n_{ij}, p_{ij}), \quad i = 1, \dots, m; j = 1, \dots, m_i. \quad (7.11)$$

Population model:

$$p_{ij}|\lambda_1, p_i \sim \text{Beta}(\lambda_1 p_i, \lambda_1 (1 - p_i)), \quad i = 1, \dots, m. \quad (7.12)$$

$$p_i|\lambda_0 \sim \text{Beta}(\lambda_0 p, \lambda_0 (1 - p)). \quad (7.13)$$

Prior distributions:

$$\lambda_1 \sim \text{Gamma}(a_1, b_1), \quad a_1 > 0, b_1 > 0. \quad (7.14)$$

$$\lambda_0 \sim \text{Gamma}(a_0, b_0), \quad a_0 > 0, b_0 > 0. \quad (7.15)$$

In the above model, m is the number of provinces, m_i is the number of RHGs within the i th province, λ_1 and λ_0 are hyperparameters, and a_1, b_1, a_0, b_0 are known constants.

Remark 1: For each RHG, the number of respondents y_{ij} is a random variable with a binomial distribution $\text{Binomial}(n_{ij}, p_{ij})$ and y_{ij} 's are independent of each other. The direct survey estimate of the response rate p_{ij} is given by

$$\hat{p}_{ij} = \frac{y_{ij}}{n_{ij}}. \quad (7.16)$$

The standard error of $\hat{p}_{ij}$ is estimated by

$$\hat{s}_{ij} = \sqrt{\frac{\hat{p}_{ij}(1 - \hat{p}_{ij})}{n_{ij}}}. \quad (7.17)$$

Note that $\hat{s}_{ij}$ is only defined for $0 < \hat{p}_{ij} < 1$. For some RHGs, the sample size n_{ij} is very small. As a result, the direct estimate $\hat{p}_{ij}$ is not reliable for these groups. We would like to obtain the posterior estimate of p_{ij} to smooth the direct estimate $\hat{p}_{ij}$. Let $Y = (y_{ij}, j = 1, \dots, m_i, i = 1, \dots, m)$. Then we are interested in evaluating $E(p_{ij}|Y)$ and the posterior variance $V(p_{ij}|Y)$.

Remark 2: Conditional on p_i , the distribution of p_{ij} is modelled by a beta distribution with parameters $\lambda_1 p_i$ and $\lambda_1(1 - p_i)$ for some $\lambda_1 > 0$. Therefore, we have

$$E(p_{ij}|\lambda_1, p_i) = p_i. \quad (7.18)$$

Thus we assume that within the i th province the response rates p_{ij} 's are random variables with conditional mean p_i , the provincial level response rate.

Remark 3: Conditional on p , the distribution of p_i among provinces is modelled by a beta distribution with parameters $\lambda_0 p$ and $\lambda_0(1 - p)$ for some $\lambda_0 > 0$. Then we have

$$E(p_i|\lambda_0, p) = p. \quad (7.19)$$

Thus the provincial response rate p_i is a random variable with mean equal to the national response rate p . The national response rate p may be calculated as

$$p = \frac{\sum_{i=1}^m \sum_{j=1}^{m_i} y_{ij}}{\sum_{i=1}^m \sum_{j=1}^{m_i} n_{ij}}. \quad (7.20)$$

p is considered as known in the model, since p is fairly stable at the national level. Note that p is based on the overall sample sizes and overall number of respondents for the whole nation.

Remark 4: The compound beta-binomial model makes it possible for two levels of smoothing, that is, the RHG level and province level. We can estimate $E(p_{ij}|Y)$ and $E(p_i|Y)$ at the same time. The direct estimate of p_i at the provincial level is given by

$$\hat{p}_i = \frac{\sum_{j=1}^{m_i} y_{ij}}{\sum_{j=1}^{m_i} n_{ij}}. \quad (7.21)$$

However, since the provinces of Canada vary largely in size and population structure, $\hat{p}_i$ can be smoothed by $E(p_i|Y)$ through the compound beta-binomial model, even though our main interest is to smooth $\hat{p}_{ij}$'s for the RHGs within provinces.

7.3.3 Gibbs sampling inference

We are interested in evaluating the posterior distributions of p_{ij} and p_i , in particular, the posterior means and posterior variances of p_{ij} and p_i . Direct evaluation involves high-dimensional integration and is not computationally feasible. Instead, we use the Gibbs sampling method. The full conditional distributions for the Gibbs sampler are:

- $[p_{ij}|Y, p_i, p, \lambda_1, \lambda_0] \propto p_{ij}^{y_{ij} + \lambda_1 p_i - 1} (1 - p_{ij})^{n_{ij} - y_{ij} + \lambda_1 (1 - p_i) - 1},$

that is, $[p_{ij}|Y, p_i, p, \lambda_1, \lambda_0]$ is a beta distribution with parameters $y_{ij} + \lambda_1 p_i$ and $n_{ij} - y_{ij} + \lambda_1(1 - p_i)$.

- $[p_i|Y, p_{ij}, p, \lambda_1, \lambda_0] \propto \frac{(\prod_{j=1}^{m_i} p_{ij})^{\lambda_1 p_i} (\prod_{j=1}^{m_i} (1 - p_{ij}))^{\lambda_1 (1 - p_i)}}{[\Gamma(\lambda_1 p_i) \Gamma(\lambda_1 (1 - p_i))]^{m_i}} p_i^{\lambda_0 p} (1 - p_i)^{\lambda_0 (1 - p)},$
- $[\lambda_1|Y, p_{ij}, p_i, p, \lambda_0] \propto \prod_{i=1}^m \left(\frac{\Gamma(\lambda_1)}{\Gamma(\lambda_1 p_i) \Gamma(\lambda_1 (1 - p_i))} \right)^{m_i} [\prod_{j=1}^{m_i} (p_{ij})^{p_i} (1 - p_{ij})^{1 - p_i}]^{\lambda_1} \lambda_1^{a_1 - 1} e^{-b_1 \lambda_1},$
- $[\lambda_0|Y, p_{ij}, p_i, p, \lambda_1] \propto \left(\frac{\Gamma(\lambda_0)}{\Gamma(\lambda_0 p) \Gamma(\lambda_0 (1 - p))} \right)^m (\prod_{i=1}^m p_i^p (1 - p_i)^{1 - p})^{\lambda_0} \lambda_0^{a_0 - 1} e^{-b_0 \lambda_0}.$

The full conditional distribution of p_{ij} is a beta distribution, thus updating p_{ij} is simple. However, Metropolis-Hastings algorithm is needed to generate samples from the conditional distributions of p_i , λ_1 and λ_0 . Since the conditional distribution of p_{ij} has a closed form, the Rao-Blackwellized estimators (Gelfand and Smith 1990, 1991) of $E(p_{ij}|Y)$ and $V(p_{ij}|Y)$ can be constructed. Suppose a sample of size G is generated from the Gibbs sampler. Noting that

$$E(p_{ij}|Y, p_i, p, \lambda_1, \lambda_0) = \frac{y_{ij} + \lambda_1 p_i}{n_{ij} + \lambda_1} \quad (7.22)$$

and

$$V(p_{ij}|Y, p_i, p, \lambda_1, \lambda_0) = \frac{(y_{ij} + \lambda_1 p_i)(n_{ij} - y_{ij} + \lambda_1(1 - p_i))}{(n_{ij} + \lambda_1)^2 (n_{ij} + \lambda_1 + 1)}, \quad (7.23)$$

it follows from $E(p_{ij}|Y) = E(E(p_{ij}|Y, p_i, p, \lambda_1, \lambda_0))$ and equation (7.22) that the Rao-Blackwellized estimator of the posterior mean of p_{ij} is given by

$$\hat{p}_{ij}^{RB} = \frac{1}{G} \sum_{k=1}^G \frac{y_{ij} + \lambda_1^{(k)} p_i^{(k)}}{n_{ij} + \lambda_1^{(k)}}. \quad (7.24)$$

To estimate the posterior variance of p_{ij} , noting that

$$V(p_{ij}|Y) = E(V(p_{ij}|Y, p_i, p, \lambda_1, \lambda_0)) + V(E(p_{ij}|Y, p_i, p, \lambda_1, \lambda_0))$$

and equations (7.22) and (7.23), we obtain the Rao-blackwellized estimator of $V(p_{ij}|Y)$ as

$$\begin{aligned}\hat{V}_{p_{ij}}^{RB} &= \frac{1}{G} \sum_{k=1}^G \frac{(y_{ij} + \lambda_1^{(k)} p_i^{(k)})(n_{ij} - y_{ij} + \lambda_1^{(k)}(1 - p_i^{(k)}))}{(n_{ij} + \lambda_1^{(k)})^2(n_{ij} + \lambda_1^{(k)} + 1)} \\ &+ \frac{1}{G} \sum_{k=1}^G \frac{(y_{ij} + \lambda_1^{(k)} p_i^{(k)})^2}{(n_{ij} + \lambda_1^{(k)})^2} - \left(\frac{1}{G} \sum_{k=1}^G \frac{y_{ij} + \lambda_1^{(k)} p_i^{(k)}}{n_{ij} + \lambda_1^{(k)}} \right)^2.\end{aligned}\quad (7.25)$$

The Rao-Blackwellized estimators can substantially reduce the simulation errors as shown in Chapter 4. Thus the beta-binomial model enables us to find stable and closed form estimators for the posterior mean and posterior variance of p_{ij} .

7.3.4 Data analysis

In our data analysis, we use 1997 HRRS data and there are 538 RHGs across Canada. Each province has different number of RHGs and the number varies largely among provinces, for example, Ontario has 153 RHGs and PEI has only 8 RHGs. Sample size for each RHG differs substantially, varying from 1 to 201. Our goal is to smooth the direct survey estimate of the response rate, particularly for the RHGs with small sample size. By using (7.20), we calculated the national response rate $p = 0.891$. Prior distributions for the hyperparameters λ_1 and λ_0 are set as $\text{Gamma}(1, 0.001)$, which is equivalent to an exponential distribution with mean 1000. We fit the compound beta-binomial model via Gibbs sampling using the BUGS 0.6 program (Spiegelhalter, Thomas, Best and Gilks, 1997).

Within each province, the HB estimators $\hat{p}_{ij}^{RB}$ smooth the direct estimators $\hat{p}_{ij}$, especially for the extreme large and small ones. Figure 7.1 displays the direct survey estimates and HB estimates of the response rates of the RHGs in Newfoundland

and the corresponding standard errors. It is clear from Figure 7.1 that the HB method has smoothed the direct estimates, and moved the RHG level response rate toward the posterior provincial level response rate. There is a substantial reduction in the standard errors, particularly for the RHGs with small sample sizes. For example, for the RHG 17 in Newfoundland, the sample size is only 4 and the direct estimate $\hat{p}_{ij} = 0.75$ with standard error equal to 0.2165, whereas the HB estimate $\hat{p}_{ij}^{RB} = 0.898$ with standard error equal to 0.0302. Figure 7.2 shows a similar result of smoothing the response rates in P.E.I. and a similar reduction in standard errors. These figures display a similar pattern and the result of smoothing at the RHG level is evident. Tables 7.1 and 7.2 present the direct estimates and the HB estimates of the response rates and their corresponding standard errors in Newfoundland and P.E.I., respectively. The results for the other provinces are given in You and Reiss (1999).

For the RHGs with $y_{ij} = n_{ij}$, we get $\hat{p}_{ij} = 1$. Thus we cannot use (7.17) to estimate the standard errors of these $\hat{p}_{ij}$. Typically for these RHGs, the sample size n_{ij} 's are relatively very small. However, by using the HB approach, we can obtain the smoothed estimate $\hat{p}_{ij}^{RB}$ and the corresponding posterior variance as the measure of uncertainty associated with $\hat{p}_{ij}^{RB}$. Overall, the HB method improves on the direct estimates at the RHG level considerably.

At the province level, Figure 7.3 shows the direct estimates and the HB estimates of the provincial level response rates and the national response rate. We note from Figure 7.3 that the provincial response rates have consistently moved towards the national response rate of 89.1%; provinces close to the national rate, such as

Quebec, Ontario and Manitoba move very little, while the extreme provinces such as Newfoundland and P.E.I. move the most. Table 7.3 shows the direct estimates, the HB estimates and the corresponding posterior standard errors for all provinces.

7.3.5 Remarks

In this section, we presented a hierarchical Bayes compound beta-binomial model to smooth the direct survey estimates of the response rates of the RHGs in the HRRS. The proposed compound beta-binomial model can smooth both the RHG level and province level response rates at the same time. Also the beta-binomial model leads to closed form full conditional distributions for p_{ij} , which enables us to obtain the Rao-Blackwellized estimators for the posterior mean and posterior variance of p_{ij} . If auxiliary variables related to the response rate are available at the RHG level, probit or logistic random effect regression models may be used to find the HB estimates of p_{ij} 's. However, with the probit or logistic model, we cannot get closed form full conditionals for p_{ij} . Nevertheless, it might be interesting to compare the results under different models and make appropriate model comparisons.

Figure 7.1: Estimates and Standard Errors (Newfoundland)

Figure 7.2: Estimates and Standard Errors (P.E.I.)

Figure 7.3: Provincial Level Response Rate Estimates

Newfoundland						
RHG	y_{ij}	n_{ij}	$\hat{p}_{ij}$	$\hat{p}_{ij}^{RB}$	$\hat{s}_{ij}$	$\hat{s}_{ij}^{RB}$
1	50	55	0.909	0.907	0.0388	0.0219
2	57	60	0.950	0.920	0.0281	0.0212
3	59	65	0.908	0.906	0.0358	0.0224
4	53	59	0.898	0.902	0.0394	0.0234
5	45	47	0.957	0.920	0.0296	0.0218
6	36	40	0.900	0.904	0.0474	0.0250
7	37	46	0.804	0.876	0.0585	0.0262
8	37	41	0.902	0.905	0.0464	0.0245
9	37	38	0.974	0.923	0.0258	0.0235
10	24	27	0.889	0.902	0.0605	0.0250
11	35	37	0.946	0.917	0.0372	0.0231
12	30	32	0.938	0.911	0.0426	0.0239
13	96	100	0.960	0.930	0.0196	0.0179
14	71	76	0.934	0.916	0.0285	0.0206
15	93	99	0.939	0.921	0.0241	0.0193
16	97	102	0.951	0.927	0.0214	0.0186
17	3	4	0.750	0.898	0.2165	0.0302
18	20	21	0.952	0.912	0.0466	0.0255
19	18	19	0.947	0.911	0.0514	0.0261
20	18	19	0.947	0.912	0.0514	0.0252
21	7	8	0.875	0.903	0.1169	0.0276

Table 7.1: Results of RHGs in Newfoundland

P.E.I.						
RHG	y_{ij}	n_{ij}	$\hat{p}_{ij}$	$\hat{p}_{ij}^{RB}$	$\hat{s}_{ij}$	$\hat{s}_{ij}^{RB}$
22	55	61	0.902	0.896	0.0381	0.0244
23	45	52	0.865	0.885	0.0474	0.0269
24	43	53	0.811	0.867	0.0538	0.0275
25	53	59	0.898	0.895	0.0394	0.0244
26	103	115	0.896	0.894	0.0285	0.0206
27	87	91	0.956	0.923	0.0215	0.0191
28	99	108	0.917	0.906	0.0265	0.0197
29	110	119	0.924	0.910	0.0243	0.0198

Table 7.2: Results of RHGs in P.E.I.

Province	No. of RHGs	y_i	n_i	$\hat{p}_i$	$\hat{p}_i^{HB}$	$\hat{s}_i^{HB}$
NFL	21	923	995	0.927	0.905	0.0085
PEI	8	595	658	0.904	0.894	0.0086
NS	49	1476	1629	0.906	0.900	0.0059
NBK	39	1290	1435	0.898	0.892	0.0077
QUB	95	3566	4013	0.889	0.889	0.0058
ONT	153	5792	6514	0.889	0.891	0.0043
MAN	39	1504	1678	0.896	0.896	0.0070
SKW	40	1630	1856	0.878	0.887	0.0067
ALT	34	1556	1756	0.886	0.889	0.0076
BC	60	1740	1991	0.874	0.880	0.0063

Table 7.3: Results for All Provinces

Chapter 8

Small Area Estimation Using Cross-sectional and Time Series Data

8.1 Introduction

Most of the research on small area estimation has focused on cross-sectional data at a given point in time. However, for repeated surveys, cross-sectional and time series models can be used to develop efficient estimators for small area characteristics by combining the direct survey estimates over time. Rao and Yu (1994) proposed a combined cross-sectional and time series model involving autocorrelated random effects and sampling errors with an arbitrary covariance matrix over time. Using this model, they obtained EBLUP estimators of small area means and associated standard errors for the current time period. Pfeiffermann and Burck (1990) and Singh, Mantel, and Thomas (1991), among others, have also considered cross-sectional and time series models for small area estimation using the Kalman filter, but they assumed specific models for the sampling error over time (e.g., independent sampling

errors). As a result, their approach cannot be used with an arbitrary covariance matrix for sampling errors. Recently, Ghosh et al. (1996) proposed a hierarchical Bayes multivariate time series model to provide the median income estimates for the U.S. states. Datta et al. (1998) applied the hierarchical Bayes approach to a cross-sectional and time series model involving a random walk model for the time component and obtained the posterior estimates of unemployment rates for the U.S. states. Like Rao and Yu (1994), Ghosh et al. (1996) and Datta et al. (1998) all assumed a known but arbitrary covariance matrix for sampling errors.

In this Chapter, we consider a univariate cross-sectional and time series model. This model is similar to the model proposed by Rao and Yu (1994) but with a random walk model for the time components instead of a autoregressive model. Following Rao and Yu (1994), we obtain EBLUP estimators of small area means and second order approximations to the MSE of the EBLUP estimators and estimators of MSE. We also apply the HB approach to this model and obtain the posterior mean and posterior variance of the small area means.

8.2 Cross-sectional and Time Series Model

Let y_{it} denote the direct survey estimate of θ_{it} , the i th small area mean at time t , $i = 1, \dots, m$; $t = 1, \dots, T$, where m is the number of small areas and T is the current time. We assume that

$$y_{it} = \theta_{it} + e_{it}, \quad (8.1)$$

where the e_{it} 's are sampling errors and the small area mean θ_{it} is modelled as

$$\theta_{it} = x'_{it}\beta + v_i + u_{it}, \quad (8.2)$$

$$u_{it} = u_{i,t-1} + \varepsilon_{it}, \quad (8.3)$$

where $x_{it} = (x_{it1}, \dots, x_{itp})'$ is the vector of auxiliary variables for the i th area at time t ; β is a vector of regression parameters of length p ; v_i is a random area effect with $v_i \stackrel{iid}{\sim} N(0, \sigma_v^2)$; for a given area i , u_{it} 's follow a random walk process over time period $t = 1, \dots, T$, $\varepsilon_{it} \stackrel{iid}{\sim} N(0, \sigma_\varepsilon^2)$. Thus $cov(u_{it}, u_{is}) = \min(t, s)\sigma_\varepsilon^2$. Let $e_i = (e_{i1}, \dots, e_{iT})'$ be the vector the sampling errors for the i th area, then we assume that e_i follows a multivariate normal distribution with mean 0 and known variance-covariance matrix Σ_i , i.e., $e_i \sim N(0, \Sigma_i)$; $\{v_i\}$, $\{\varepsilon_{it}\}$ and $\{e_i\}$ are mutually independent. Specification of Σ_i may not be easy in practice. A direct estimator of Σ_i is likely to be unstable, particularly for the areas with small sample sizes. Therefore, smoothed estimator of Σ_i is often used and used as a proxy for Σ_i .

Arranging the data $\{y_{it}\}$ as a vector $y = (y'_1, \dots, y'_m)'$ with $y_i = (y_{i1}, \dots, y_{iT})'$, we can write models (8.1) and (8.2) in matrix form as

$$y_i = X_i\beta + 1_T v_i + u_i + e_i, \quad i = 1, \dots, m, \quad (8.4)$$

or

$$y = X\beta + Zv + u + e \quad (8.5)$$

with

$$X = (X'_1, \dots, X'_m)', \quad X'_i = (x_{i1}, \dots, x_{iT}),$$

$$Z = I_m \otimes 1_T, \quad v = (v_1, \dots, v_m),$$

$$u = (u'_1, \dots, u'_m)', \quad u'_i = (u_{i1}, \dots, u_{iT}),$$

$$e = (e'_1, \dots, e'_m)', \quad e'_i = (e_{i1}, \dots, e_{iT}),$$

where I_m is the identity matrix of order m , 1_T is a $T \times 1$ vector of 1's, and $\otimes$ denotes the direct Kroneckor product. Further,

$$E(v) = 0, \quad \text{cov}(v) = \sigma_v^2 I_m, \quad (8.6)$$

$$E(u) = 0, \quad \text{cov}(u) = \sigma_\varepsilon^2 I_m \otimes \Gamma \equiv \sigma_\varepsilon^2 R, \quad (8.7)$$

$$E(e) = 0, \quad \text{cov}(e) = \Sigma = \text{blockdiag}(\Sigma_1, \dots, \Sigma_m), \quad (8.8)$$

and v , u and e are mutually independent, where Γ is a $T \times T$ matrix with elements equal to $\min(t, s)$, $t, s = 1, \dots, T$. It follows from (8.5)-(8.8) that

$$\begin{aligned} \text{cov}(y) &= V = \Sigma + \sigma_\varepsilon^2 R + \sigma_v^2 Z Z' \\ &= \text{blockdiag}(V_1, \dots, V_m) \end{aligned} \quad (8.9)$$

with $V_i = \Sigma_i + \sigma_\varepsilon^2 \Gamma + \sigma_v^2 J_T$, where $J_T = 1_T 1_T'$.

8.3 BLUP Estimator

We are interested in estimating the current small area mean $\theta_{iT} = x'_{iT} \beta + v_i + u_{iT}$, which is a special case of the linear combination $\tau = l' \beta + l'_1 v + l'_2 u$, where $l = x_{iT}$, l_1 is a $m \times 1$ vector with 1 in the i th position and 0 elsewhere, and l_2 is a $mT \times 1$ vector with 1 in the iT th position and 0 elsewhere. Noting that model (8.5) is a special case of the general mixed linear model given in Chapter 1, the BLUP estimator of θ_{iT} can be obtained from Henderson's general results (Henderson 1975; also see Chapter 1) for arbitrary linear combinations τ .

Assuming that σ_ε^2 and σ_v^2 are known, the BLUP estimator of τ is given by

$$\tilde{\tau} = l' \tilde{\beta} + (\sigma_v^2 l'_1 Z' + \sigma_\varepsilon^2 l'_2 R) V^{-1} (y - X \tilde{\beta}), \quad (8.10)$$

where $\tilde{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}y$ is the generalized least squares estimator of β . Using the special structures of l_1, l_2, Z, R and V , it is easy to find the BLUP estimator of θ_{iT} as

$$\begin{aligned}\tilde{\theta}_{iT} &= x'_{iT}\tilde{\beta} + (\sigma_v^2 \mathbf{1}_T + \sigma_\varepsilon^2 \gamma'_T)' V_i^{-1} (y_i - X_i \tilde{\beta}) \\ &\equiv h(\sigma_\varepsilon^2, \sigma_v^2, y),\end{aligned}\tag{8.11}$$

where $V_i = \Sigma_i + \sigma_\varepsilon^2 \Gamma + \sigma_v^2 J_T$, and $\gamma_T = (1, 2, \dots, T)$ is the T th row of matrix Γ . Thus $\tilde{\theta}_{iT}$ is a function depending on the unknown variance components σ_ε^2 and σ_v^2 .

8.4 EBLUP Estimator

The variance components σ_ε^2 and σ_v^2 are usually unknown in practice. In this section, following Rao and Yu (1994), we use the method of moments to obtain consistent estimators of σ_ε^2 and σ_v^2 , and then obtain the EBLUP estimator for θ_{iT} .

8.4.1 Estimation of σ_ε^2

We first obtain an unbiased estimator of σ_ε^2 . For this purpose, we transform models (8.1) and (8.2) to eliminate the random effect v_i . First transform y_i to $z_i = Py_i$ such that the covariance matrix of Pu_i is $\sigma_\varepsilon^2 I_T$, i.e., $\Gamma = P^{-1}(P')^{-1}$. The transformed model is given by

$$z_i = PX_i\beta + f v_i + P(u_i + e_i), i = 1, \dots, m.\tag{8.12}$$

where f is a $T \times 1$ vector given as $f = P \cdot \mathbf{1}_T = (f_1, \dots, f_T)'$. Next we transform z_i to w_i as $w_i = (I_T - D)z_i$, where $D = (ff')/c$ with $c = f'f = \mathbf{1}_T P' P \mathbf{1}_T$. This leads

to the following reduced model:

$$w_i = H_i\beta + e_i^*, \quad i = 1, \dots, m, \quad (8.13)$$

where $H_i = (I_T - D)PX_i$ and $e_i^* = (I_T - D)P(u_i + e_i)$.

Since

$$\text{Cov}(e_i^*) = (I_T - D)(\sigma_\varepsilon^2 I_T + P\Sigma_i P')(I_T - D)' \quad (8.14)$$

does not involve σ_v^2 , we can estimate σ_ε^2 through the reduced model (8.13) using the residual sum of squares. Let $w = (w_1', \dots, w_m')'$, $H = (H_1', \dots, H_m')'$, and $\hat{e}'\hat{e}$ be the residual sum of squares obtained by regressing w on H using ordinary least squares. Then we have

$$\begin{aligned} \hat{e}'\hat{e} &= w'(I - H(H'H)^{-1}H')w \\ &= (e^{*'} + \beta'H')(I - H(H'H)^{-1}H')(H\beta + e^*) \\ &= e^{*'}(I - H(H'H)^{-1}H')e^*. \end{aligned} \quad (8.15)$$

An unbiased estimator of σ_ε^2 is thus obtained by

$$\begin{aligned} \bar{\sigma}_\varepsilon^2 &= (\hat{e}'\hat{e} - \text{tr}[(\text{blockdiag}_i(I_T - D) - H(H'H)^{-1}H') \\ &\quad \times \text{blockdiag}_i(P\Sigma_i P')])(m(T - 1) - r(H))^{-1}, \end{aligned} \quad (8.16)$$

where A^- is a generalized inverse of A . The unbiasedness of $\bar{\sigma}_\varepsilon^2$ follows by noting (8.15) and the following lemma:

Lemma 8.1: If y is a random p -dimension multivariate with mean 0 and covariance matrix Ω , then $E(y'Gy) = \text{tr}(G\Omega)$ for any symmetric matrix G .

Since $\bar{\sigma}_\varepsilon^2$ can take negative values, we truncate it at zero and use

$$\hat{\sigma}_\varepsilon^2 = \max\{0, \bar{\sigma}_\varepsilon^2\}. \quad (8.17)$$

The truncated estimator $\hat{\sigma}_\varepsilon^2$ is no longer unbiased, but it is asymptotically consistent with bounded (small or moderate) number of time points, T , as $m \rightarrow \infty$.

8.4.2 Estimation of σ_v^2

Turning to the estimation of σ_v^2 , we transform (8.12) by changing z_i to $c^{-1/2}f'z_i$ such that $u_i^* = c^{-1/2}f'Pu_i$ has mean 0 and variance σ_ε^2 . The transformed model is given by

$$c^{-1/2}f'z_i = c^{-1/2}f'PX_i\beta + c^{1/2}v_i + u_i^* + c^{-1/2}f'Pe_i \quad (8.18)$$

with error variance $c\sigma_v^2 + \sigma_\varepsilon^2 + c^{-1}f'P\Sigma_iP'f$. Let $\hat{u}'\hat{u}$ be the residual sum of squares obtained by regressing $c^{-1/2}f'z_i$ on $c^{-1/2}f'PX_i$ using ordinary least squares. An unbiased estimator of σ_v^2 is then given by

$$\begin{aligned} \bar{\sigma}_v^2 = & c^{-1}(m - r(F))^{-1}[\hat{u}'\hat{u} \\ & - \text{tr}((I_m - F(F'F)^{-1}F')\text{diag}_i(c^{-1}f'P\Sigma_iP'f))] - c^{-1}\bar{\sigma}_\varepsilon^2, \end{aligned} \quad (8.19)$$

where $F = (X_1'P'f, \dots, X_m'P'f)'$. The unbiasedness of $\bar{\sigma}_v^2$ follows by noting that

$$E(\hat{u}'\hat{u}) = (c\sigma_v^2 + \sigma_\varepsilon^2)(m - r(F)) + \text{tr}[(I_m - F(F'F)^{-1}F')\text{diag}_i(c^{-1}f'P\Sigma_iP'f)].$$

Also, noting that $\bar{\sigma}_v^2$ can take negative values, we truncate it at zero and use

$$\hat{\sigma}_v^2 = \max\{0, \bar{\sigma}_v^2\}. \quad (8.20)$$

The truncated estimator $\hat{\sigma}_v^2$ is not unbiased, but it is also asymptotically consistent as $m \rightarrow \infty$.

8.4.3 EBLUP estimator of θ_{iT}

An EBLUP estimator of θ_{iT} is now obtained by substituting $\hat{\sigma}_\varepsilon^2$ and $\hat{\sigma}_v^2$ for σ_ε^2 and σ_v^2 respectively in (8.11):

$$\hat{\theta}_{iT} = h(\hat{\sigma}_\varepsilon^2, \hat{\sigma}_v^2, y). \quad (8.21)$$

The EBLUP estimator $\hat{\theta}_{iT}$ remains unbiased, noting that $\hat{\sigma}_\varepsilon^2$ and $\hat{\sigma}_v^2$ are even functions of y and translation invariant, i.e., they remain unchanged when y is changed to $-y$ or $y - Xa$ for all y and a (Kacker and Harville, 1984).

It is not necessary to assume normality of the error terms in the models (8.1) to (8.3). Only symmetric distributions are needed. However, we need the normality assumption to derive the MSE of $\hat{\theta}_{iT}$ and its estimator, correct to a second-order approximation.

8.5 MSE Estimation

We first obtain a second-order approximation to the MSE of the EBLUP estimator $\hat{\theta}_{iT}$, in the sense that the neglected terms are of order $o(m^{-1})$ for large m . Using this approximation, an estimator of $MSE(\hat{\theta}_{iT})$, correct to the same order of approximation, is then obtained.

8.5.1 Second-order approximation to the MSE

Following Kackar and harville (1984), we have, under normality of the errors v_i , u_{it} and e_{it} ,

$$MSE(\hat{\theta}_{iT}) = MSE(\tilde{\theta}_{iT}) + E(\tilde{\theta}_{iT} - \hat{\theta}_{iT})^2, \quad (8.22)$$

where $MSE(\tilde{\theta}_{iT}) = E(\tilde{\theta}_{iT} - \theta_{iT})^2$. Further, using Henderson's general result given in Chapter 1, an exact expression for $MSE(\tilde{\theta}_{iT})$ is given by

$$MSE(\tilde{\theta}_{iT}) = g_{1iT}(\sigma_\varepsilon^2, \sigma_v^2) + g_{2iT}(\sigma_\varepsilon^2, \sigma_v^2), \quad (8.23)$$

where

$$g_{1iT}(\sigma_\varepsilon^2, \sigma_v^2) = \sigma_v^2 + \sigma_\varepsilon^2 T - (\sigma_v^2 \mathbf{1}_T + \sigma_\varepsilon^2 \gamma'_T)' V_i^{-1} (\sigma_v^2 \mathbf{1}_T + \sigma_\varepsilon^2 \gamma'_T), \quad (8.24)$$

and

$$g_{2iT}(\sigma_\varepsilon^2, \sigma_v^2) = (x_{iT} - X_i' V_i^{-1} (\sigma_v^2 \mathbf{1}_T + \sigma_\varepsilon^2 \gamma'_T))' (X' V^{-1} X)^{-1} (x_{iT} - X_i' V_i^{-1} (\sigma_v^2 \mathbf{1}_T + \sigma_\varepsilon^2 \gamma'_T)). \quad (8.25)$$

The second term $g_{2iT}(\sigma_\varepsilon^2, \sigma_v^2)$ in (8.23), due to estimating β , is of order $O(m^{-1})$, while the first term $g_{1iT}(\sigma_\varepsilon^2, \sigma_v^2)$ is $O(1)$.

It remains to evaluate the second term $E(\tilde{\theta}_{iT} - \hat{\theta}_{iT})^2$ in (8.22). Following Kackar and Harville (1984), we propose a Taylor approximation to this term:

$$E(\tilde{\theta}_{iT} - \hat{\theta}_{iT})^2 \approx E\left[\frac{\partial \tilde{\theta}_{iT}}{\partial \sigma_\varepsilon^2} (\bar{\sigma}_\varepsilon^2 - \sigma_\varepsilon^2) + \frac{\partial \tilde{\theta}_{iT}}{\partial \sigma_v^2} (\bar{\sigma}_v^2 - \sigma_v^2)\right]^2.$$

Following Prasad and Rao (1990), a further approximation is obtained as

$$E(\tilde{\theta}_{iT} - \hat{\theta}_{iT})^2 \approx \text{tr}[\Delta' V \Delta \Sigma^*] \equiv g_{3iT}(\sigma_\varepsilon^2, \sigma_v^2), \quad (8.26)$$

where Σ^* is the 2×2 covariance matrix of the unbiased estimators $\bar{\sigma}_\varepsilon^2$ and $\bar{\sigma}_v^2$. Further $\Delta = (\partial b / \partial \sigma_\varepsilon^2, \partial b / \partial \sigma_v^2)$ where b is a $mT \times 1$ vector given as

$$b' = (0, \dots, 0, b'_i, 0, \dots, 0)$$

with $b'_i = (\sigma_v^2 \mathbf{1}_T + \sigma_\varepsilon^2 \gamma'_T)' V_i^{-1}$. Then we have $\Delta' V \Delta = \Delta'_i V_i \Delta_i$, where $\Delta_i = (\partial b_i / \partial \sigma_\varepsilon^2, \partial b_i / \partial \sigma_v^2)$. Calculating the derivatives $\partial b_i / \partial \sigma_\varepsilon^2$ and $\partial b_i / \partial \sigma_v^2$, we obtain,

after simplification, $\Delta_i' V_i \Delta_i = A$, where A is a 2×2 symmetric matrix with diagonal elements

$$\begin{aligned} a_{11} &= [\gamma_T' - \Gamma V_i^{-1}(\sigma_v^2 \mathbf{1}_T + \sigma_\varepsilon^2 \gamma_T')] V_i^{-1} [\gamma_T' - \Gamma V_i^{-1}(\sigma_v^2 \mathbf{1}_T + \sigma_\varepsilon^2 \gamma_T')], \\ a_{22} &= [\mathbf{1}_T - J_T V_i^{-1}(\sigma_v^2 \mathbf{1}_T + \sigma_\varepsilon^2 \gamma_T')] V_i^{-1} [\mathbf{1}_T - J_T V_i^{-1}(\sigma_v^2 \mathbf{1}_T + \sigma_\varepsilon^2 \gamma_T')], \end{aligned}$$

and off-diagonal elements

$$a_{12} = a_{21} = [\gamma_T' - \Gamma V_i^{-1}(\sigma_v^2 \mathbf{1}_T + \sigma_\varepsilon^2 \gamma_T')] V_i^{-1} [\mathbf{1}_T - J_T V_i^{-1}(\sigma_v^2 \mathbf{1}_T + \sigma_\varepsilon^2 \gamma_T')].$$

The term $g_{3iT}(\sigma_\varepsilon^2, \sigma_v^2)$ is of the same order as $g_{2iT}(\sigma_\varepsilon^2, \sigma_v^2)$. Combining (8.23) and (8.26), we get a second-order approximation to $MSE(\hat{\theta}_{iT})$ as

$$MSE(\hat{\theta}_{iT}) = g_{1iT}(\sigma_\varepsilon^2, \sigma_v^2) + g_{2iT}(\sigma_\varepsilon^2, \sigma_v^2) + g_{3iT}(\sigma_\varepsilon^2, \sigma_v^2). \quad (8.27)$$

The neglected terms in the approximation (8.27) are of lower order, $o(m^{-1})$, for large m . A rigorous proof of this assertion is quite involved, but follows along the lines of Prasad and Rao (1990).

It remains to obtain the covariance matrix Σ^* . For this purpose we rewrite $\bar{\sigma}_\varepsilon^2$ and $\bar{\sigma}_v^2$ as

$$\bar{\sigma}_\varepsilon^2 = [(m-1)T - \text{rank}(H)]^{-1} a' C_1 a + \text{const.} \quad (8.28)$$

$$\begin{aligned} \bar{\sigma}_v^2 &= c^{-1} [m - \text{rank}(F)]^{-1} a' C_2 a \\ &\quad - c^{-1} [(m-1)T - \text{rank}(H)]^{-1} a' C_1 a + \text{const.} \end{aligned} \quad (8.29)$$

where $a = Zv + u + e \sim N(0, V)$,

$$C_1 = C'[I_n - CX(X'CC'X)^{-1}X'C']C$$

with $C = \text{blockdiag}_i[(I_T - D)P]$ and

$$C_2 = C^{**'}[I_m - C^{**'}X(X'C^{**'}C^*X)^{-1}X'C^{**'}]C^*$$

with $C^* = \text{blockdiag}_i(c_i^{-1/2}f'P)$. We can now evaluate the elements of Σ^* using (8.28) and (8.29) in the following well-known lemma on the covariance of two quadratic forms of normally distributed variables.

Lemma 8.2: If $y \sim N(0, \Omega)$, then $\text{Cov}(y'G_1y, y'G_2y) = 2\text{tr}(G_2\Omega G_1\Omega)$, where G_1 and G_2 are two symmetric matrices.

8.5.2 Second-order approximation to estimator of MSE

Following Prasad and Rao (1990), it can be shown that

$$E(g_{1iT}(\hat{\sigma}_\varepsilon^2, \hat{\sigma}_v^2)) \approx g_{1iT}(\sigma_\varepsilon^2, \sigma_v^2) - g_{3iT}(\sigma_\varepsilon^2, \sigma_v^2), \quad (8.30)$$

$$E(g_{2iT}(\hat{\sigma}_\varepsilon^2, \hat{\sigma}_v^2)) \approx g_{2iT}(\sigma_\varepsilon^2, \sigma_v^2), \quad (8.31)$$

and

$$E(g_{3iT}(\hat{\sigma}_\varepsilon^2, \hat{\sigma}_v^2)) \approx g_{3iT}(\sigma_\varepsilon^2, \sigma_v^2). \quad (8.32)$$

The neglected terms in the approximations (8.30) - (8.32) are of lower order, $o(m^{-1})$.

A rigorous proof of this assertion is quite involved, but essentially involves showing that $E(o_p(m^{-1})) = o(m^{-1})$.

It now follows from (8.30) - (8.32) and (8.27) that an estimator of $MSE(\hat{\theta}_{iT})$ with expectation correct to $O(m^{-1})$ is given by

$$mse(\hat{\theta}_{iT}) = g_{1iT}(\hat{\sigma}_\varepsilon^2, \hat{\sigma}_v^2) + g_{2iT}(\hat{\sigma}_\varepsilon^2, \hat{\sigma}_v^2) + 2g_{3iT}(\hat{\sigma}_\varepsilon^2, \hat{\sigma}_v^2). \quad (8.33)$$

A naive estimator of $MSE(\hat{\theta}_{iT})$, which ignores the uncertainty in the estimators $\hat{\sigma}_\varepsilon^2$ and $\hat{\sigma}_v^2$, is given by

$$mse_N(\hat{\theta}_{iT}) = g_{1iT}(\hat{\sigma}_\varepsilon^2, \hat{\sigma}_v^2) + g_{2iT}(\hat{\sigma}_\varepsilon^2, \hat{\sigma}_v^2). \quad (8.34)$$

The naive MSE estimator (8.34) is computationally simpler than (8.33), since it does not require $g_{3iT}(\hat{\sigma}_\varepsilon^2, \hat{\sigma}_v^2)$, which involves the estimation of covariance matrix of the variance components $\hat{\sigma}_\varepsilon^2$ and $\hat{\sigma}_v^2$. However, it can lead to significant underestimation of the true MSE.

8.6 Hierarchical Bayes Approach

In this section, we apply the hierarchical Bayes approach to the cross-sectional and time series model given by (8.1) - (8.3). Estimators of the posterior mean and posterior variance of the interested small area mean are obtained using the Gibbs sampling method.

8.6.1 The hierarchical Bayes model

We consider the following hierarchical Bayes model:

(i) Conditional on $\theta_i = (\theta_{i1}, \dots, \theta_{iT})'$,

$$y_i | \theta_i \sim N(\theta_i, \Sigma_i), \quad i = 1, \dots, m;$$

(ii) Conditional on β, u_{it} and σ_v^2 ,

$$\theta_{it} | \beta, u_{it}, \sigma_v^2 \sim N(x'_{it}\beta + u_{it}, \sigma_v^2), \quad i = 1, \dots, m, t = 1, \dots, T;$$

(iii) Conditional on $u_{i,t-1}$ and σ_ε^2 ,

$$u_{it}|u_{i,t-1}, \sigma_\varepsilon^2 \sim N(u_{i,t-1}, \sigma_\varepsilon^2), \quad i = 1, \dots, m, t = 2, \dots, T;$$

(iv) Marginal prior distributions of β , σ_v^2 and σ_ε^2 :

$$\beta \sim \text{uniform}(R^p);$$

$$\sigma_v^2 \sim IG(a_1, b_1) \propto (\sigma_v^2)^{-(a_1+1)} e^{-\frac{b_1}{\sigma_v^2}};$$

$$\sigma_\varepsilon^2 \sim IG(a_2, b_2) \propto (\sigma_\varepsilon^2)^{-(a_2+1)} e^{-\frac{b_2}{\sigma_\varepsilon^2}}.$$

We are interested in estimating $E(\theta_i|Y)$ and $V(\theta_i|Y)$, in particular, the current time posterior mean and posterior variance $E(\theta_{iT}|Y)$ and $V(\theta_{iT}|Y)$.

8.6.2 Gibbs sampling inference

To implement the Gibbs sampler, we need to generate samples from the following full conditional distributions:

- $\beta|Y, \sigma_v^2, \sigma_\varepsilon^2, u, \theta \sim N(\sigma_v^2(X'X)^{-1}X'(\theta - u), \sigma_v^2(X'X)^{-1});$
- $\sigma_v^2|Y, \beta, \sigma_\varepsilon^2, u, \theta \sim IG(a_1 + \frac{mT}{2}, b_1 + \frac{1}{2} \sum_{i=1}^m \sum_{t=1}^T (\theta_{it} - x'_{it}\beta - u_{it})^2);$
- $\sigma_\varepsilon^2|Y, \beta, \sigma_v^2, u, \theta \sim IG(a_2 + \frac{m(T-1)}{2}, b_2 + \frac{1}{2} \sum_{i=1}^m \sum_{t=2}^T (u_{it} - u_{i,t-1})^2);$
- for $i = 1, \dots, m,$

$$u_{i1}|Y, \beta, \sigma_v^2, \sigma_\varepsilon^2, u_{i2}, \theta \sim N((\frac{1}{\sigma_v^2} + \frac{1}{\sigma_\varepsilon^2})^{-1}(\frac{\theta_{i1} - x'_{i1}\beta}{\sigma_v^2} + \frac{u_{i2}}{\sigma_\varepsilon^2}), (\frac{1}{\sigma_v^2} + \frac{1}{\sigma_\varepsilon^2})^{-1});$$
- for $i = 1, \dots, m,$

$$u_{iT}|Y, \beta, \sigma_v^2, \sigma_\varepsilon^2, u_{i,T-1}, \theta \sim N((\frac{1}{\sigma_v^2} + \frac{1}{\sigma_\varepsilon^2})^{-1}(\frac{\theta_{iT} - x'_{iT}\beta}{\sigma_v^2} + \frac{u_{i,T-1}}{\sigma_\varepsilon^2}), (\frac{1}{\sigma_v^2} + \frac{1}{\sigma_\varepsilon^2})^{-1});$$

- for $i = 1, \dots, m$, and $2 \leq t \leq T - 1$,

$$u_{it}|Y, \beta, \sigma_v^2, \sigma_\varepsilon^2, u_{i,t-1}, u_{i,t+1}, \theta \sim N((\frac{1}{\sigma_v^2} + \frac{2}{\sigma_\varepsilon^2})^{-1}(\frac{\theta_{it} - \frac{\sigma_\varepsilon^2}{\sigma_v^2}\beta}{\sigma_\varepsilon^2} + \frac{u_{i,t-1} + u_{i,t+1}}{\sigma_\varepsilon^2}), (\frac{1}{\sigma_v^2} + \frac{2}{\sigma_\varepsilon^2})^{-1});$$

- for $i = 1, \dots, m$,

$$\theta_i|Y, \beta, \sigma_v^2, \sigma_\varepsilon^2, u \sim N((\sigma_v^{-2}I_T + \Sigma_i^{-1})^{-1}(\Sigma_i^{-1}y_i + \sigma_v^{-2}(X_i\beta + u)), (\sigma_v^{-2}I_T + \Sigma_i^{-1})^{-1}).$$

Using the Gibbs sampler, the Rao-Blackwellized estimators of the posterior mean $E(\theta_i|Y)$ and posterior variance $V(\theta_i|Y)$ of the θ_i are given respectively by

$$\hat{\theta}_i = \frac{1}{G} \sum_{k=1}^G (\sigma_v^{-2(k)}I_T + \Sigma_i^{-1})^{-1}(\Sigma_i^{-1}y_i + \sigma_v^{-2(k)}(X_i\beta^{(k)} + u_i^{(k)})) \quad (8.35)$$

and

$$\begin{aligned} \hat{V}(\theta_i|Y) &= \frac{1}{G} \sum_{k=1}^G (\sigma_v^{-2(k)}I_T + \Sigma_i^{-1})^{-1} \\ &\quad + \frac{1}{G} \sum_{k=1}^G (\sigma_v^{-2(k)}I_T + \Sigma_i^{-1})^{-1}(\Sigma_i^{-1}y_i + \sigma_v^{-2(k)}(X_i\beta^{(k)} + u_i^{(k)})) \\ &\quad \times (\Sigma_i^{-1}y_i + \sigma_v^{-2(k)}(X_i\beta^{(k)} + u_i^{(k)}))^T (\sigma_v^{-2(k)}I_T + \Sigma_i^{-1})^{-1} \\ &\quad - \frac{1}{G^2} [\sum_{k=1}^G (\sigma_v^{-2(k)}I_T + \Sigma_i^{-1})^{-1}(\Sigma_i^{-1}y_i + \sigma_v^{-2(k)}(X_i\beta^{(k)} + u_i^{(k)}))] \\ &\quad \times [\sum_{k=1}^G (\sigma_v^{-2(k)}I_T + \Sigma_i^{-1})^{-1}(\Sigma_i^{-1}y_i + \sigma_v^{-2(k)}(X_i\beta^{(k)} + u_i^{(k)}))]^T. \end{aligned} \quad (8.36)$$

where $\{\beta^{(k)}, \sigma_v^{2(k)}, u_i^{(k)}, k = 1, \dots, G\}$ is a sample of size G generated from the Gibbs sampler. Thus by using the Gibbs sampling approach, we can not only estimate the current time small area mean θ_{iT} , but also find the estimators of small area means $\theta_{it}, t = 1, \dots, T$, simultaneously for each area i . The posterior variance matrix estimator $\hat{V}(\theta_i|Y)$ also provides the estimate of covariance $cov(\theta_{it}, \theta_{is})$ for $t, s = 1, \dots, T$.

Chapter 9

Future Research

In this thesis, we have applied the hierarchical Bayes approach to various small area models and have obtained the posterior estimators for small area characteristics under these models. The MCMC sampling method enables us to consider various complex statistical models for small area estimation. For future research, we propose the following directions:

In Chapter 3, a Fay-Herriot type model was used for census undercoverage rate estimation. But that model has some limitations as noted in Chapter 3. We can, however, consider the following hierarchical non-linear models:

Model 1:

- $y_i = \theta_i + e_i$, $e_i \sim N(0, \psi_i^2)$, where $\psi_i^2 = \frac{(1-\theta_i)^4}{C_i^2} \xi_i^2$;
- $\log(\theta_i) = x_i^T \beta + v_i$, $v_i \sim N(0, \sigma_v^2)$.

Model 2:

- $\hat{U}_i = U_i + \epsilon_i$, $\epsilon_i \sim N(0, \xi_i^2)$, where ξ_i^2 known;
- $\log(\frac{U_i}{U_i + C_i}) = x_i^T \beta + v_i$, $v_i \sim N(0, \sigma_v^2)$.

In Model 1, sampling variance ψ_i^2 is a function of the unknown mean θ_i . For the proposed new models, we can use the same hierarchical Bayes approach as that used in Chapter 6. Model 2 has the advantage that the sampling variance is known, unlike in the case of Model 1. It would be very interesting to compare the non-linear modelling with the linear modelling used in Chapter 3.

In Chapter 5, we considered pseudo hierarchical Bayes estimation for small area means by combining small area models and direct survey estimators. Prasad and Rao (1999) proposed a pseudo EBLUP approach for the same problem. It will be of interest to compare these two approaches via a simulation study and on real survey data. Also, recently Folsom et al. (1999) proposed a survey weighted Gibbs algorithm based on survey weighted pseudo log-likelihood function for binary variables in small area estimation. It is of great interest to consider their approach in a linear model case and compare their approach with our pseudo hierarchical Bayes approach.

In Chapter 6, we considered a unit-level log linear model for small area estimation, where we assumed that the sampling model for the observed units y_{ij} is linear and the linking model for small area means is log linear. Also the auxiliary variable x_i is at the area-level. We intend to consider an extension to more general nonlinear models with auxiliary information available at the unit-level.

In Chapter 7, we considered binary data analysis in small area estimation. We used a probit model to predict the small area means. We intend to consider to bring sampling weights into the model and get model-assisted design-based predictors for small area means. The idea of survey weighted pseudo log-likelihood proposed by

Folsom et al. (1999) could be used. For the proposed compound beta-binomial model, no auxiliary variable is used in the model. However, if auxiliary information is available at certain stage in a multi-stage survey design, a mixture model involving beta-binomial and logistic or probit models could be considered. Also further investigation should be made on the effects of using smoothed Bayes estimators for the nonresponse adjustment.

It would be of interest to apply the cross-sectional and time series model of Chapter 8 to a real survey problem and compare the results of EBLUP and HB approaches. One interesting application would be in the Canadian Labour Force Survey (LFS). The LFS produces monthly estimates of unemployment rates at Canada level and provincial level. It also releases unemployment estimates for subprovincial areas such as Economic Regions (ERs) across Canada. Estimates for Canada and provinces are fairly reliable with a CV of about 2% for Canada and 4% to 7% for provinces. However, for subprovincial areas, CVs are usually between 15% to 25%. Thus, for some subprovincial areas, the estimates are not reliable since the overall sample size is determined to provide accurate estimates for the nation and all provinces. The cross-sectional and time-series models can be used to borrow strength across regions and time periods to produce reliable subprovincial unemployment estimates.

Bibliography

- [1] Albert, J.H. and Chib, S. (1993) Bayesian analysis of binary and ploychotomous response data. *Journal of the American Statistical Association*, 88, 669-679.
- [2] Arora, A. and Lahiri, P. (1995) On the superiority of the Bayesian methods over the BLUP in samll area estimation problems. *Proceedings of the Survey Methods Section*, Statistical Society of Canada Annual Meeting, Montreal, Quebec, July, 1995.
- [3] Battese. G.E., Harter, R.M. and Fuller, W.A. (1988) An error components model for prediction of county crop area using survey and satellite data. *Journal of the American Statistical Association*, 83, 28-36.
- [4] Bell, W. (1998) Models for county and state poverty estimates. In *Small-Area Estimation of School-Age Children in Poverty: Interim Report 2-Evaluation of Revised 1993 County Estimates for Title I Allocations*, Edited by C.F. Citro, M.L. Cohen, and G. Kalton, National Research Council, Washington D.C.: National Academy Press.
- [5] Berger, J.O. (1985) *Statistical Decision Theory and Bayesian Analysis*, 2nd edition. New York: Springer-Verlag.

- [6] Best, N., Cowles, M.K., and Vines, k. (1996) *CODA, Convergence diagnosis and output analysis software for Gibbs sampling output, Version 0.30*. MRC Biostatistics Unit, Institute of Public Health, Robinson Way, cambridge CB2 2SR.
- [7] Box, G.E.P. and Tiao, G.C. (1973) *Bayesian Inference In Statistical Analysis*, Addison-Wesley.
- [8] Brooks, S.P. (1998) Markov chain Monte Carlo method and its application. *The Statistician*, 47, 69-100.
- [9] Chip, S. and Carlin, B.P. (1997) On MCMC sampling in hierarchical longitudinal models. Manuscript.
- [10] Chip, S. and Greenberg, E. (1995) Understanding the Metropolis-Hastings Algorithm. *The American Statistician*, 49, 327-335.
- [11] Czado, C. (1994) Bayesian inference of binary regression models with parametric link. *Journal of Statistical Planning and Inference*, 41, 121-140.
- [12] Czado, C. and You, Y. (1993) Inference for binary regression models with parametric link: a comparison. *Technical Report*, York University, Toronto, Canada.
- [13] Datta, G.S. and Ghosh, M. (1991) Bayesian prediction in linear models: Applications to small area estimation. *The Annals of Statistics*, 19, 1748-1770.
- [14] Datta, G.S., Ghosh, M., Huang, E.T., Isaki, C.T., Schultz, L.K. and Tsay, J.H. (1992) Hierarchical and empirical Bayes methods for adjustment of census

- undercount: The 1988 Missouri dress rehearsal data. *Survey Methodology*, 18, 95-108.
- [15] Datta, G.S. and Lahiri, P. (1995) Robust hierarchical Bayes estimation of small area characteristics in the presence of covariates and outliers. *Journal of Multivariate Analysis*, 54, 310-328.
- [16] Datta, G.S., Lahiri, P., Maiti, T. and Lu, K.L. (1999) Hierarchical Bayes estimation of unemployment rates for the U.S. states. Manuscript.
- [17] Dagpunar, J. (1988) *Principles of Random Variate Generation*. Clarendon Press, Oxford.
- [18] Dick, P. (1995) Modelling net undercoverage in the 1991 Canadian census. *Survey Methodology*, 21, 45-54.
- [19] Dick, P. and You, Y. (1997) Bayes and census undercoverages. *Proceedings of the Survey Methods Section*, 57-66, Statistical Society of Canada Annual Meetings, Fredericton, New Brunswick, June, 1997.
- [20] Dick, P. and You, Y. (1997) A hierarchical Bayes analysis of Census undercoverage. *Proceedings of Symposium 97 New Directions in Surveys and Censuses*, 101-105, Statistics Canada, Ottawa.
- [21] Ericken, E.P. and Kadane, J.B. (1985) Estimating the population in a census year (with discussion). *Journal of the American Statistical Association*, 80, 98-131.

- [22] Farrell, P.J. (1995) A hierarchical Bayes approach for estimating small area proportions. *Proceedings of the Survey Methods Section*, 63-68, Statistical Society of Canada Annual Meetings, Montreal, Quebec, July, 1995.
- [23] Farrell, P.J., MacGibbon, B. and Tomberlin, T.J. (1997) Empirical Bayes small area estimation using logistic regression models and summary statistics. *Journal of Business and Economic Statistics*, 15, 101-108.
- [24] Fay, R.E. and Herriot, R.A. (1979) Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.
- [25] Folsom, R., Shah, B. and Vaish, A. (1999) Substance abuse in States: model-based estimates from the 1994-1996 National Household Surveys on Drug Abuse methodology report. A draft report.
- [26] Geman, S. and Geman, D. (1984) Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721-741.
- [27] Gelman, A. and Rubin, D.B. (1992) Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 457-472.
- [28] Gelfand, A.E. (1995) Model determination using sampling-based methods. In *Markov Chain Monte Carlo in Practice* (eds W.R. Gilks, S. Richardson, and D.J. Spiegelhalter), pp. 145-161. London: Chapman & Hall.

- [29] Gelfand, A.E. and Smith, A.F.M. (1990) Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398-409.
- [30] Gelfand, A.E. and Smith, A.F.M. (1991) Gibbs sampling for marginal posterior expectations. *Communications In Statistics - Theory and Methods*, 20, 1747-1766.
- [31] Ghosh, M., Nangia, N. and Kim, D.H. (1996) Estimation of median income of four-person families: a Bayesian time series approach. *Journal of the American Statistical Association*, 91, 1423-1431.
- [32] Ghosh, M., Natarajan, K., Stroud, T.W.F. and Carlin, B.P. (1998) Generalized linear models for small area estimation. *Journal of the American Statistical Association*, 93, 273-282.
- [33] Ghosh, M. and Rao, J.N.K. (1994) Small area estimation: An appraisal (with discussion). *Statistical Science*, 9, 55-93.
- [34] Gilks, W.R. and Wild, P. (1992) Adaptive rejection sampling for Gibbs sampling. *Journal of Applied Statistics*, 41, 337-348.
- [35] Gilks, W.R., Best, N.G. and Tan, K.K.C. (1995) Adaptive rejection Metropolis sampling within Gibbs sampling. *Journal of Applied Statistics*, 44, 455-472.
- [36] Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. (1996) *Markov Chain Monte Carlo in Practice*, London: Chapman & Hall.
- [37] Hastings, W.K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 87, 97-109.

- [38] He, Z. and Sun, D. (1998) Bayes estimation of success rates for small areas. Manuscript.
- [39] Henderson, C.R. (1975) Best linear unbiased estimation and prediction under a selection model. *Biometrics*, 31, 423-447.
- [40] Hobert, J.P. and Cassella, G. (1996) The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association*, 91, 1461-1473.
- [41] Holt, D. and Moura, F. (1993) Small area estimation using multi-level models. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 1, 21-30.
- [42] Holt, D. and Moura, F. (1999) Small area estimation using survey data and multi-level models. *Survey Methodology*, in press.
- [43] Jiang, J. and Lahiri, P. (1998) Empirical best prediction for small area inference with binary data. Manuscript.
- [44] Jiang, J., Lahiri, P. and Wan, S.M. (1998) Jackknifing the mean squared error of empirical best predictor. Manuscript.
- [45] Kacker, R.N. and Harville, D.A. (1984) Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistical Association*, 79, 853-862.

- [46] 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.
- [47] Kott, P. (1989) Robust small domain estimation using random effects modelling. *Survey Methodology*, 23, 81-89.
- [48] Lahiri, P. and Rao, J.N.K. (1995) Robust estimation of mean squared error of small area estimators. *Journal of the American Statistical Association*, 90, 758-766.
- [49] Lindley, D.V. (1983) Comment on parametric empirical Bayes inference: theory and applications (by C.N. Morris). *Journal of the American Statistical Association*, 78, 61-62.
- [50] Lindley, D.V. and Smith, A.F.M. (1972) Bayes estimates for the linear model. *Journal of the Royal Statistical Society, B*, 34, 1-41.
- [51] MacGibbon, B. and Tomberlin, T.J. (1989) Small area estimates of proportions via empirical Bayes techniques. *Survey Methodology*, 15, 237-252.
- [52] Maiti, T. (1998) Hierarchical Bayes estimation of mortality rates for disease mapping. *Journal of Statistical Planning and Inference*, 69, 339-348.
- [53] Malec, D., Sedransk, J., Moriarity, C.L., and Leclerc, F.B. (1997) Small area inference for binary variables in the national health interview survey. *Journal of the American Statistical Association*, 92, 815-826.

- [54] Metropolis, N., Rosenbluth, A.W., Rosenbluth, m.N., Teller, A.H. and Teller, E. (1953) Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21, 1087-1091.
- [55] Morris, C. (1983) Parametric empirical Bayes inference: Theory and applications (with discussion). *Journal of the American Statistical Association*, 78, 47-65.
- [56] Pfeffermann, D. and Burck, L. (1990) Robust small area estimation combining time series and cross-sectional data. *Survey Methodology*, 16, 217-237.
- [57] Prasad, N.G.N. and Rao, J.N.K. (1990) The estimation of mean squared error of small area estimators. *Journal of the American Statistical Association*, 85, 163-171.
- [58] Prasad, N.G.N. and Rao, J.N.K. (1999) On robust samll area estimation using a simple random effects model. *Survey Methodology*, in press.
- [59] Rao, J.N.K. (1997) Developments in sample survey theory: an appraisal. *The Canadian Journal of Statistics*, 25, 1-21.
- [60] Rao, J.N.K. (1999) Some current trends in sample survey theory and methods. Manuscript.
- [61] Rao, J.N.K. (1999) Small area estimation: Updates with appraisal. Manuscript.
- [62] Rao, J.N.K. and Yu, M. (1994) Small area estimation by combining time series and cross-sectional data. *The Canadian Journal of Statistics*, 22, 511-528.

- [63] Rivest, L.P. (1995) A composite estimator for provincial undercoverage in the Canadian census. *Proceedings of the Survey Method Section*, Statistical Society of Canada Annual Meeting, Montreal, Quebec, July, 1995.
- [64] Robert, C.P. (1994) *The Bayesian Choice: A Decision-Theoretic Motivation*. New York: Springer-Verlag.
- [65] Robert, C.P. (1995) Simulation of truncated normal variables. *Statistics and Computing*, 5, 121-125.
- [66] Robinson, G.K. (1991) That BLUP is a good thing: The estimation of random effects (with discussion). *Statistical Science*, 6, 15-51.
- [67] Rosner, B. (1989) Multivariate methods for clustered binary data with more than one level of nesting. *Journal of the American Statistical Association*, 84, 373-380.
- [68] Royce, D. (1992) A comparison of some estimators of a set of population totals. *Survey Methodology*, 18, 109-125.
- [69] Singh, A.C., Mantel, H. and Thomas, B.W. (1991) Time series generalizations of Fay-Herriot estimator for small areas. *Proceedings of Survey Research Methods Section*, American Statistical Association, 455-460.
- [70] Smith, A.F.M. and Roberts, G.O. (1993) Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society*, B, 55, 3-23.

- [71] Spiegelhalter, D., Thomas, A., Best, N., and Gilks, W. (1996) *BUGS 0.5, Bayesian inference using Gibbs sampling manual*. MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge CB2 2SR.
- [72] Stroud, T.W.F. (1994) Bayesian analysis of binary survey data. *The Canadian Journal of Statistics*, 22, 33-45.
- [73] Stukel, D.M. and Rao, J.N.K. (1999) On small-area estimation under two-fold nested error regression models. *Journal of Statistical Planning and Inference*, 78, 131-147.
- [74] Tanner, M. and Wong, W. (1987) The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association*, 82, 528-550.
- [75] Tierney, L. (1994) Markov chains for exploring posterior distributions (with discussion). *Annals of Statistics*, 22, 1701-1762.
- [76] You, Y (1996) *Hierarchical Bayes estimation of Canadian Census undercoverage via Gibbs sampler*. Statistics Canada Internship Project, Ottawa.
- [77] You, Y (1997) *Hierarchical Bayes models for Census undercoverage estimation*. Internal Statistics Canada Report, Ottawa.
- [78] You, Y and Rao, J.N.K. (1997) Hierarchical Bayes small area estimation using multi-level models. *Proceedings of Symposium 97 New Directions in Surveys and Censuses*, 278-281, Statistics Canada, Ottawa.

- [79] You, Y and Rao, J.N.K. (1999) Hierarchical Bayes estimation of small area means using multi-level models. Invited paper, *Conference Proceedings, IASS Satellite Conference on Small Area Estimation*, 171-185, Riga, Latvia.
- [80] You, Y and Rao, J.N.K. (1999) Pseudo hierarchical Bayes small area estimation using sampling weights. *Proceedings of the Survey Methods Section*, Statistical Society of Canada Annual Meetings, Regina, Saskatchewan, June, 1999.
- [81] You, Y and Reiss, P. (1999) Hierarchical Bayes estimation of response rates for an expenditure survey. *Proceedings of the Survey Methods Section*, Statistical Society of Canada Annual Meetings, Regina, Saskatchewan, June, 1999.
- [82] Zaslavavsky. A.M. (1993) Combining census, Dual-System, and evaluation study data to estimate population shares. *Journal of the American Statistical Association*, 88, 1092-1105.
- [83] Zeger, S.L. and Karim, M.R. (1991) Generalized linear models with random effects: A Gibbs's sampling approach. *Journal of the American Statistical Association*, 86, 79-86.