Adaptive Hierarchical Bayes Estimation of Small-Area Proportions

Benmei Liu[1]
[1]National Cancer Institute, 6116 executive Blvd., Suite 504, Rockville, MD 20852

Abstract
Unit level logistic regression models with mixed effects have been used for estimating small-area proportions in the literature. Normality is commonly assumed for the random effects. Nonetheless, real data often show significant departures from normality assumptions of the random effects. To reduce the risk of model misspecification, we propose an adaptive hierarchical Bayes estimation approach in which the distribution of the random effect is chosen adaptively from the exponential power class of probability distributions. The richness of the exponential power class ensures the robustness of our hierarchical Bayes approach against departure from normality. We demonstrate the robustness of our proposed model using simulated data and real data.

Keywords*:* Proportions, random effects, Exponential Power distribution

1. Introduction

During the past three decades, the demand for statistical estimates for small areas using large-scale survey data has increased dramatically in many different application areas including income and poverty, education, health, and agriculture. For budget restrains, the same survey data, originally designed to provide statistically reliable, design based estimates of characteristics of interest for a high level of aggregation (e.g., national level, large geographic domains such as region), also is used to generate estimates at a lower level (e.g., states, counties, etc.). In absence of adequate direct information from sample survey data, model-based methods that use mixed models to combine information from the survey data and external sources such as Census and administrative records have been proposed in small area estimation. Both empirical best prediction (EBP) and hierarchical Bayesian (HB) approaches have been used for inference using area level or unit level mixed models. Rao (2003) gave a whole range of review on both approaches. Jiang and Lahiri (2006a) provided extensive review on recent development of the EBP approach in small area estimation. In this research, we consider the situation when the characteristics of interest are binary (i.e., 0 or 1) and the estimates to be produced are small proportions at small area level.

To estimate small-area proportions, logistic regression models with small area specific effects are commonly used. For example, in order to estimate true small-area proportions $P_i = \Sigma_{k=1}^{N_i} y_{ik} / N_i$, MacGibbon and Tomberlin (1989) considered the following models:

$$\begin{aligned} & y_{ik} \mid p_{ik} \overset{ind}{\sim} Bernoulli(p_{ik}), \\ & \log it(p_{ik}) = \log(p_{ik} / (1-p_{ik})) = \mathbf{x}_{ik}'\boldsymbol{\beta} + v_i;\ v_i \overset{iid}{\sim} N(0, \sigma_v^2), \end{aligned} \tag{1.1}$$

where p_{ik} denotes the probability of a response for the k th unit in the i th area, and y_{ik} and $\mathbf{x}_{ik}$, $k = 1,...,N_i;\ i = 1,...,m$, are unit-specific response of the characteristic and covariates respectively. The model-based estimator of P_i was obtained using $\hat{p}_i = \Sigma_{k=1}^{N_i} \hat{p}_{ik} / N_i$, where $\hat{p}_{ik}$ is obtained from (1.1) by estimating $\boldsymbol{\beta}$ and the

realization of v_i through empirical Bayes method. The applications of similar models can also be found in Farrell, MacGibbon, and Tomberlin (1997a and b), among others. Malec et al. (1997) considered a different logistic regression model with random regression coefficients. Suppose each individual in the population is assigned to one of K mutually exclusive and exhaustive classes based on the individual's socioeconomic/demographic status. The binary response y_{ijl} for individual l $(l = 1,...,N_{ij})$ in class j in cluster i is assumed independent Bernoulli with common probability p_{ij}. To make inference on a finite population proportion for a specified small area and subgroup $P = \sum_{i\in I}\sum_{j\in K}\sum_{l=1}^{N_{ij}} y_{ijl} \Big/ \sum_{i\in I}\sum_{j\in K} N_{ij}$, where I is the collection of clusters that define the small area and K is the collection of classes that defines the subpopulation, the following models are assumed:

$$\begin{aligned} & y_{ijl} \mid p_{ij} \overset{ind}{\sim} Bernoulli(p_{ij}); \\ & \log it(p_{ij}) = \log(p_{ij} / (1 - p_{ij})) = \mathbf{x}'_j \boldsymbol{\beta}_i;\ \ \boldsymbol{\beta}_i = \mathbf{Z}_i \alpha + v_i;\ \ v_i \overset{iid}{\sim} N(0, \boldsymbol{\Sigma}_v); \end{aligned} \tag{1.2}$$

where $\mathbf{x}_j$ is class-specific covariate vector and $\mathbf{Z}_i$ is a $p \times q$ area level covariate matrix. They used HB approach based on model (1.2) to estimate the proportion (overall and socioeconomic/demographic group) of persons in a state or substate who have visited a physician in the past year, using the data from the National Health Interview Survey.

Logistic regression mixed models typically assume normality for the area-level random effects (e.g., see model 1.1 and 1.2). The wide use of the normality assumption can be attributed to its conceptual and computational simplicity as well as its popularity in standard data analysis. Nevertheless, we would expect that certain type of measurements would not be normally distributed. For example, leptokurtic (kurtosis>0) distributions and platykurtic distributions (kurtosis<0) for individual errors can occur (e.g., see Chapter 3 of Box and Tiao, 1973). For cases where the assumption of normality is not tenable, more flexible models can be adopted to accommodate non-normality. However, the literature in small area estimation on this aspect is not rich.

Farrell, MacGibbon and Tomberlin (1994) considered the EBP approach for protecting against outlying parameters. Using a simple random-effect model which is a special case of model (1.1), Farrell et al. compared the effects of step-function priors with those of the normal and Laplace priors for the random effects. They found that as the tails of the prior become heavier, the Laplace distribution is the most appropriate prior. For skewed prior distributions, the use of a step-function prior was recommended. To the best of our knowledge, this is the only research paper addressing non-normality problem in the application of logistic regression models for estimating small-area proportions in the small area estimation literature.

To accommodate non-normality patterns such as kurtosis for the random effects, we propose a robust unit level mixed model by assuming a class of distributions which includes normal for the random effects under complex sampling design. We make inference on small-area proportions using survey data under a stratified simple random sample (SRS) design for this research, where the small areas are the design strata.

We briefly review the exponential power distribution in Section 2. In Section 3, we present a motivating example for this study. In Section 4, we propose a robust unit level model for survey data drawn from a finite population using a stratified SRS design to accommodate kurtosis problems. In Section 5, we illustrate some Bayesian inference procedures based on the proposed model. In Section 6, we evaluate the proposed model

by comparing it with the normal model using some purely simulated data and several real datasets. This chapter finishes with some concluding remarks in Section 7.

2. Exponential Power Distribution

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

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

where $\mu \in R,\ \sigma \in \mathrm{R}^+,\ \varphi \in (0,1],\quad c_0 = \Gamma(3\varphi)/\Gamma(\varphi)$, $c_1 = \sqrt{c_0}\,/\left(2\varphi\Gamma(\varphi)\right)$. The parameters $\mu,\ \sigma,\ \varphi$ are location, scale and shape (kurtosis) parameters respectively. This parameterization is preferred to the more popular one proposed by Box and Tiao (1973) because it implies $E(X) = \mu$ and $Var(X) = \sigma^2$, a property that can be very useful in modeling. This family includes a range of symmetric distributions that change gradually from the uniform ($\varphi \to 0$), through short-tailed distributions (platikurtik) to the normal ($\varphi = 0.5$), then through distributions with longer-than-normal tails (leptokurtic) to the double exponential shape ($\varphi = 1$). Figure 1 illustrates the EP distributions with common mean $\mu = 0$ and standard deviation $\sigma = 1$ for six fix values of φ. The excess of kurtosis is $\gamma = \dfrac{\Gamma(\varphi)\Gamma(5\varphi)}{\Gamma^2(3\varphi)} - 3$.

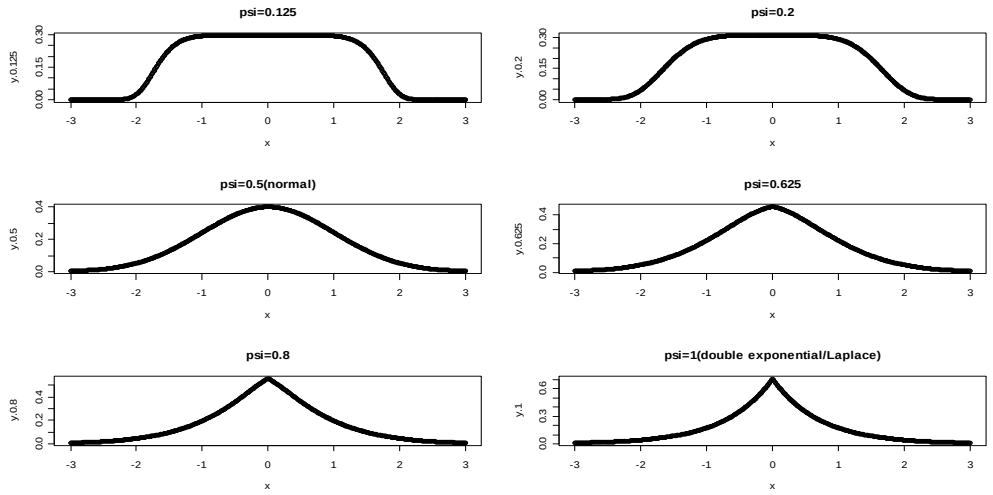

Figure 1: EP density plot with $\mu = 0,\ \sigma = 1$ for different φ

The exponential power distribution family can be very useful as a model in Monte Carlo robustness studies because it can attain a broad range of kurtosis values and include three well-known symmetric distributions as special cases. Box and Tiao (1973) used this family extensively as an alternative to the normal distribution for statistical modeling and also as a tool to study Bayesian robustness. In all the examples they studied, they found that the inferences on population mean could differ substantially as the kurtosis parameter changes. Hogg (1974) discussed the exponential power distribution family with

$0.5 \le \varphi \le 1$ in relation to adaptive estimators of location. Prescott (1978) studied the asymptotic properties of the φ-trimmed means and other adaptive trimmed means from this family of distributions. For normal location problem, Choy and Smith (1997a) used the Laplace approximation method for integrals to approximate the posterior moments for the leptokurtic class of the exponential power distribution family and found that this subclass of distributions robustifies the estimation procedure by downweighting the influence of outlying observations. For random effects models, Choy and Smith (1997b) made use of the scale mixture of normal representation of the leptokurtic density function for use in conjunction with Markov Chain Monte Carlo methods.

3. Motivating example – Low Birthweight Data

Birthweight is one of the most accessible and most understood variables in epidemiology. A baby's weight at birth is a strong indicator not only of a birth mother's health and nutritional status but also a newborn's chances for survival, growth, long-term health and psychosocial development. Babies born weighing less than 5 pounds, 8 ounces (2,500 grams) are considered low birthweight. In contrast, the average newborn weighs about 7 pounds. Over 7 percent of all newborn babies in the United States have low birthweight. Low-birthweight babies are at increased risk of serious health problems as newborns, lasting disabilities and even death. The overall rate of these very small babies in the United States is increasing (http://www.healthsystem.virginia.edu/uvahealth/peds_hrnewborn/lbw.cfm).

For evaluation purpose, we studied the estimation of state level low birthweight using samples drawn from a known population. We treated the 2002 Natality public-use data file as our study population. The file included all births occurring within the United States in 2002. Details about the births recorded in the National Vital Statistics System are given at the website for the National Center for Health Statistics (http://www.cdc.gov/nchs/births.htm). The finite population studied comprised 4,024,378 records of live births with birth weights reported in the 50 U.S. states plus the District of Columbia. The parameter of interest was the state level low birthweight rates P_i, $i = 1,...,51$.

We wanted to fit model (1.1) assuming that all the individuals in state i have a common probability P_i. Two auxiliary variables were selected to fit the model after stepwise model selection process. Let $(y_{ik}, x_{ik1}, x_{ik2})$ denote the indicator of low birthweight and two binary auxiliary variables (percentage of mother's age less than 15 and percentage of being the first child in the family) associated with the k th baby in the i th state $(k = 1,...,N_i;\ i = 1,...,51)$, and let (P_i, x_{i1}, x_{i2}) denote the corresponding state level mean. We obtained P_i, x_{1i}, x_{2i} from the population data and then fitted the following logistic regression model:

$$\text{logit}(P_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + v_i, \qquad (3.1)$$

where $v_i \overset{iid}{\sim} N(0, \sigma^2)$, $i = 1,...,51$.

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

i) Kolmogorov-Smirnov (K-S) normality test;
ii) normal Quantile-Quantile (Q-Q) plot;
iii) Bayesian method.

The p -value from the K-S test is 0.043, which indicates that v_i are not normal at significant level $\alpha = 0.05$. The left panel of Figure 3-1 displays the normal Q-Q plot for v_i. The plot indicates that the underlying distribution of v_i is more like a platikurtic distribution. To verify this, we produced the descriptive statistics for v_i using SAS PROC UNIVARIATE and the results of the descriptive statistics confirmed that v_i are platikurtic.

Since SAS uses different parameterization, to estimate the kurtosis of the residuals v_i , we considered Bayesian approach. We assumed a priori independence between the components of (σ, φ) and specified the following non-informative priors: i) $\varphi \sim Unif(0, 1)$ and ii) $\sigma \sim Unif(0, K)$, K is a given large positive number. For reference on this prior assumption, we refer to Gelman (2006). We implemented model $v_i \sim EP(0, \sigma, \varphi)$ with the two prior assumptions using WinBUGS software. The posterior mean of φ is 0.2. The one-sided 95% credible interval of φ is (0, 0.473), which does not include the normal case ($\varphi = 0.5$). The posterior density plot of the kurtosis parameter φ is displayed on the left panel of Figure 3. Clearly, the posterior mode of φ is around 0.23 and the chance of covering normal case is very small. To assess the model fit, we applied the Deviance Information Criterion (DIC) (Spiegelhalter et al. 2002) criterion and compared the fit of the alternative models for different given kurtosis. Among several alternatives including the normal model, the EP model with $\varphi = 0.2$ fits the data best since it resulted in the smallest DIC. We then compared the Q-Q plot of v_i with the Q-Q plot of a random data generated from a platikurtic exponential power distribution with $\varphi = 0.2$ (see the right panel of Figure 2). The similarity between the two Q-Q plots further confirms that the underlying distribution of v_i is platikurtic.

All these analyses demonstrated that a platikurtic EP distribution (with $\varphi < 0.5$) describes the underlying distribution of residual v_i better than the normal distribution for the Natality data. Based on the prior assumption $\sigma \sim Unif(0, 100)$, We display the posterior density plot of the scale parameter σ on the right panel of Figure 2. The mode of σ is around 0.12 , that is, the mode of the variance σ^2 is around 0.01, which is very small. According to Gelman (2006), uniform prior is preferred to inverse-gamma prior for σ^2 since it is so small.

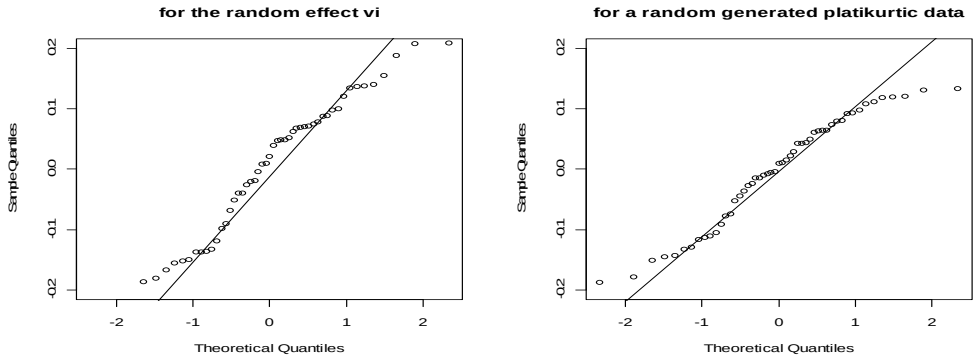

Figure 2: Normal Q-Q Plots for residual v_i and randomly generated data from platikurtic EP distribution

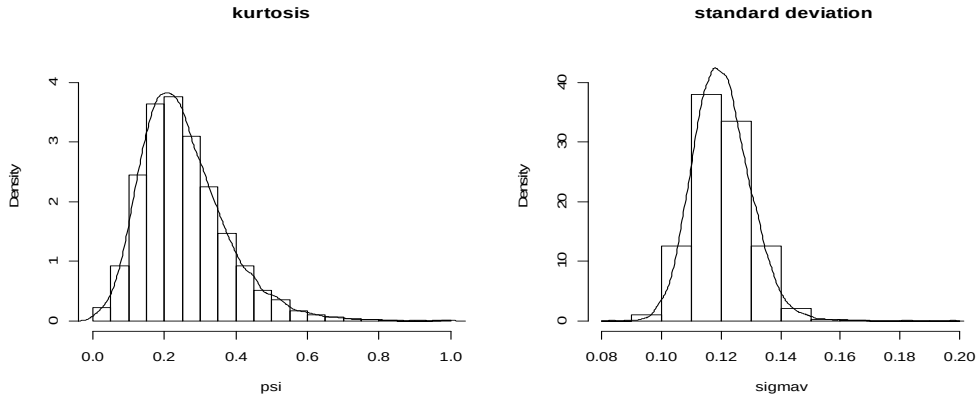

Figure 3: Posterior density of the hyperparameters φ and σ

4. Small Area Model

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

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

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

$$\text{Level 2:}\quad \text{logit}(\theta_i)=\mathbf{x}_i'\boldsymbol{\beta}+v_i,\ \ i=1,...,m \tag{4.2}$$

$$\text{where}\quad v_i \overset{iid}{\sim} N(0,\sigma^2),\ \ i=1,...,m. \tag{4.3}$$

Here θ_i s are the model parameters for the expectation of y_{ik}. For convenience, we call the model of (4.1)~(4.3) *Bernoulli-Logit-Normal* model. As demonstrated in Section 3, the normal assumption for the random effects v_i in (4.3) can not accommodate the kurtosis problem. One can possibly assume a specific distribution from the exponential power family such as Laplace (double exponential) distribution. However, there is still mis-specification risk for the distribution of the random effects. To improve robustness, instead of assuming normal or other specific non-normal distribution, we assume that the

random effects v_i follow an unspecified distribution belonging to the exponential power distribution family with two parameters:

$$v_i \overset{iid}{\sim} EP(0, \sigma, \varphi). \tag{4.4}$$

We call the proposed model (4.1)-(4.2)-(4.4) *Bernoulli-Logit-EP* model. We assume the hyperparameters σ and φ are both unknown. The strength of our proposed model is that we use a class of probability distributions instead of a specific one and the underlying model will be chosen adaptively by the data.

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

5. Bayesian Inference

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

$$\begin{aligned} P_i &= \frac{1}{N_i}\left(\Sigma_{k\in s_i} y_{ik} + \Sigma_{k\in s_i^c} y_{ik}\right) \\ &= \frac{1}{N_i}\left\{n_i p_i + \left(N_i - n_i\right) p_{ins}\right\} \\ &= f_i p_i + (1 - f_i) p_{ins}, \end{aligned} \tag{5.1}$$

where $f_i = n_i / N_i$ is the sampling rate and $1 - f_i$ is the finite population correction, p_i and p_{ins} are the area level proportions based on sampled and nonsampled units respectively. Since p_i are known given the sample, from (5.1), we can say that the prediction of P_i is equivalent to the prediction of p_{ins} given the sample.

Thus the Bayes estimator of P_i is:

$$\begin{aligned} E(P_i \mid \mathbf{y}_s) &= f_i p_i + \left(1 - f_i\right) E\left(p_{ins} \mid \mathbf{y}_s\right) \\ &= f_i p_i + \left(1 - f_i\right) \frac{1}{N_i - n_i} \Sigma_{k\in s_i^c} E\left(y_{ik} \mid \mathbf{y}_s\right) \\ &= f_i p_i + \left(1 - f_i\right) \frac{1}{N_i - n_i} \Sigma_{k\in s_i^c} E\left\{E\left(y_{ik} \mid \theta_i, \mathbf{y}_s\right) \mid \mathbf{y}_s\right\} \\ &= f_i p_i + \left(1 - f_i\right) E\left(\theta_i \mid \mathbf{y}_s\right) \equiv g_i\left[E(\theta_i \mid \mathbf{y}_s)\right], \end{aligned} \tag{5.2}$$

where $\theta_i = \dfrac{\exp(\mathbf{x}_i'\boldsymbol{\beta} + v_i)}{1 + \exp(\mathbf{x}_i'\boldsymbol{\beta} + v_i)}$. From (5.2), we can see that once $E(\theta_i \mid \mathbf{y}_s)$ is estimated, it is straightforward to estimate $E(P_i \mid \mathbf{y}_s)$ if f_i is known. We can also see that $E\left(\theta_i \mid \mathbf{y}_s\right)$ is a good approximation of $E(P_i \mid \mathbf{y}_s)$ if $f_i \approx 0$. Further, the posterior variance of P_i is given by:

$$V(P_i \mid \mathbf{y}_s) = V\left\{E(P_i \mid \mathbf{y}_s, \theta_i) \mid \mathbf{y}_s\right\} + E\left\{V(P_i \mid \mathbf{y}_s, \theta_i) \mid \mathbf{y}_s\right\}. \tag{5.3}$$

Since $E(P_i \mid \mathbf{y}_s, \theta_i) = f_i p_i + (1 - f_i)\theta_i$ and $V(P_i \mid \mathbf{y}_s, \theta_i) = \frac{1}{N_i}(1 - f_i)\theta_i(1-\theta_i)$, formula (5.3) can be further written as:

$$\begin{aligned} V(P_i \mid \mathbf{y}_s) &= (1 - f_i)^2 V(\theta_i \mid \mathbf{y}_s) + \frac{1}{N_i}(1 - f_i) E\{\theta_i(1-\theta_i) \mid \mathbf{y}_s\} \\ &= (1 - f_i)^2 V(\theta_i \mid \mathbf{y}_s) + \frac{1}{N_i}(1 - f_i)\left\{E(\theta_i \mid \mathbf{y}_s) - V(\theta_i \mid \mathbf{y}_s) - E^2(\theta_i \mid \mathbf{y}_s)\right\} \\ &= (1 - f_i)\left[\left(1 - f_i - \frac{1}{N_i}\right)V(\theta_i \mid \mathbf{y}_s) + \frac{1}{N_i} E(\theta_i \mid \mathbf{y}_s)\{1 - E(\theta_i \mid \mathbf{y}_s)\}\right] \\ &\equiv h_i\left[V(\theta_i \mid \mathbf{y}_s),\ E(\theta_i \mid \mathbf{y}_s)\right], \end{aligned} \tag{5.4}$$

Formula (5.4) indicates that $V(P_i \mid \mathbf{y}_s)$ is a linear function of $V(\theta_i \mid \mathbf{y}_s)$ and $E(\theta_i \mid \mathbf{y}_s)$. Once $V(\theta_i \mid \mathbf{y}_s)$ and $E(\theta_i \mid \mathbf{y}_s)$ are obtained, it is straightforward to obtain $V(P_i \mid \mathbf{y}_s)$ if f_i and N_i are known. We can also see from (5.4) that $V(P_i \mid \mathbf{y}_s) \approx V(\theta_i \mid \mathbf{y}_s)$ holds if $f_i \approx 0$.

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

$$P_i \mid \mathbf{y}_s \overset{ind}{\sim} N\left(E(P_i \mid \mathbf{y}_s),\ V(P_i \mid \mathbf{y}_s)\right). \tag{5.5}$$

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

In this research, we consider the case when $f_i \approx 0$ which occurs often in large-scale surveys. As we demonstrated earlier, the inference on P_i is equivalent to the inference on θ_i if $f_i \approx 0$. As a result, the Bayesian inference will be focused on the posterior distribution:

$$f(\theta_1, ..., \theta_m \mid y_s) = \int_{\boldsymbol{\beta}} \int_{\sigma} \int_{\varphi} f(\theta_1, ..., \theta_m, \boldsymbol{\beta}, \sigma, \varphi \mid y_s) d\boldsymbol{\beta} d\sigma d\varphi .$$

Base on the small area models described in Section 4, the joint posterior distribution $f(\theta_1, ..., \theta_m, \boldsymbol{\beta}, \sigma, \varphi \mid y_s)$ cannot be expressed in a single closed form, some approximation is needed. However, the joint posterior distribution can be simulated using a Markov Chain Monte Carlo (MCMC) such as Gibbs sampling or the Metropolis-Hastings algorithm. Following Malec et al. (1997), we will make inference on θ_i through HB approach and implement the proposed model using the MCMC technique. The posterior mean $E(\theta_i \mid \mathbf{y}_s)$ approximates the HB point estimate of P_i and the posterior variance of $V(\theta_i \mid \mathbf{y}_s)$ is used as a measure of variability.

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

6. Model Evaluation and Data Analysis

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

6.1 Simulated Data Analysis

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

1) When v_i are non-normal, how worse the *Bernoulli-Logit-Normal* model performs compared to the *Bernoulli-Logit-EP* model;

2) When v_i are actually normal, what is the effect for over-parameterization by the *Bernoulli-Logit-EP* model.

To generate the data, we set $n_i = 5$ and $m = 100$. We also set 4 different cases of σ and φ by varying the values:

i) $\sigma^2 = 0.01$ and 0.1;
ii) $\varphi = 0.2$ (platikutic) and 0.5 (normal).

For each of the 4 combined cases of σ and φ, we generated one sample data from the sequential models: $v_i \sim EP(0, \sigma, \varphi)$, $\text{logit}(\theta_i) = \mu + v_i$, and $y_{ij} \sim Bernoulli(P_i)$, $j = 1,...,n_i$, $i = 1,...,m$. Without loss of generality, we set $\mu = 0$. To implement the HB modeling using the sampled data, for simplicity, we assumed no auxiliary variables were available, i.e., $\mathbf{x}_i'\boldsymbol{\beta} = \mu$. we also specified the following individual prior assumptions for individual parameters: i) Flat prior for μ, i.e., $f(\mu) \propto 1$; ii) $\sigma \sim Uniform(0, L)$, and iii) $\varphi \sim Uniform(0,\ 1)$.

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

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

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

Table 1 reports the ratios of the evaluation statistics for the HB estimates based on the model *Bernoulli-Logit-Normal* over those based on the alternative model. Both evaluation statistics show consistent patterns for the HB estimates. When the random effects v_i were generated from EP distribution with $\varphi = 0.2$ (see the first two rows in the table), *Bernoulli-Logit-Normal* model produces worse results than the *Bernoulli-Logit-EP*

model. For example, the loss is over 13 percent in terms of AAD for the first case. When the random effects v_i were generated from normal distribution (see the last two rows in the table), *Bernoulli-Logit-EP* model still gives better results compared with the *Bernoulli-Logit-Normal* model. Overall, the results indicate that the effect for over-parameterization is not worrisome and the *Bernoulli-Logit-EP* model is pretty robust. The table also shows that when σ^2 is larger, the results from the two models are closer, which means that the results are less sensitive to the kurtosis measure.

Table 1: Ratios of the evaluation statistics for the two models (Normal/EP) using simulated data

Data Generate Model	AAD	AARD
$EP(\mu = 0, \sigma = 0.1, \varphi = 0.2)$	1.131	1.132
$EP(\mu = 0, \sigma = 0.33, \varphi = 0.2)$	1.029	1.027
$N\left(\mu = 0, \sigma^2 = 0.01\right)$	0.992	0.992
$N\left(\mu = 0, \sigma^2 = 0.11\right)$	0.996	0.996

6.2 Real Data Analysis

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

6.2.1 Sample Data Drwan from the 2002 Natality Population

We revisit the birthweight problem using the 2002 Natality public-use data as described in Section 3. We drew six sets of independent samples of size $n = 4,526$ using simple random sampling within states from the finite population. The state level sample sizes n_i ranged from 7 (for small states such as Vermont) to 690 (for California). The sample sizes n_i are the same as those used in Liu, Lahiri and Kalton (2008). The sampling fraction f_i varied from 0.0007 to 0.0046 which are approximated equal to zero.

In this analysis, we wanted to compare the performance of the two models: *Bernoulli-Logit-EP* and *Bernoulli-Logit-Normal*. Using each sampled data, we computed the HB estimates based on both models incorporating the two auxiliary variables considered in Section 3. The prior distributions on the hyperparameters are identical to the ones used in Section 6.1. To compare the two models, the four evaluation statistics, described in Section 6.1, are again computed for each HB estimator. Table 2 reports the ratio of the evaluation statistics for *Bernoulli-Logit-Normal* to *Bernoulli-Logit-EP*. The numbers in the table consistently show that *Bernoulli-Logit-EP* model works better than the *Bernoulli-Logit-Normal* model in terms of the four evaluation statistics. As we demonstrated in Section 3, the random effects v_i for this data set are not normal. Therefore, the analysis result in this subsection is consistent with what we found using purely simulated data in Section 6.1.

Table 2. Ratio of the four summary statistics for the HB estimates produced by *Bernoulli-Logit-Normal* over those produced by *Bernoulli-Logit-EP*

Sample	AAD	AARD
1	1.016	1.016
2	1.045	1.040
3	1.021	1.012
4	1.023	1.016
5	1.076	1.076
6	1.014	1.013

6.2.2 Baseball Data

In this subsection, we revisit the well-known baseball data given in Efron and Morris (1975). This data set has been analyzed by several researchers in the past, including Efron and Morris (1975), Gelman et al. (1995), Rao (2003), Jiang and Lahiri (2006a), among others. The data set contains the batting averages of 18 major league players through their first 45 official at bats of the 1970 season (p_i) and the true batting averages of all the 18 players for the rest of the 1970 season (p_{ins}). Efron and Morris (1975) used this data set to demonstrate the performance of their empirical Bayes and limited translation empirical Bayes estimators derived using an exchangeable prior in the presence of an outlying observation. They considered the problem of predicting the batting average for all the players for the remainder of the 1970 season based on their batting averages for the first 45 at bats. Gelman et al. (1995) provided additional data for this estimation problem and included important auxiliary data like the batting average of each player in the previous (1969) season. We consider the same estimation problem as Efron and Morris (1975) did. That is, we want to predict p_{ins} using the sampled data.

The sample size $n_i = 45$ is the number of times at bats for each player, $i = 1,...,18$. Using the baseball data, we computed the HB estimates for p_{ins} using the two models: *Bernoulli-Logit-EP* and *Bernoulli-Logit-Normal*. The previous season batting average was used as a covariate. We can prove that $E\left(p_{ins} \mid \mathbf{y}_s\right) = E\left(\theta_i \mid \mathbf{y}_s\right)$ and $V\left(p_{ins} \mid \mathbf{y}_s\right) = V\left(\theta_i \mid \mathbf{y}_s\right)$ based on the two models. Figure 4 displays the true rest 1970 season batting average (Ptrue) for each player along with the sample proportion (DirectP) and the two different HB estimators (HBEP and HBNorm) in the increasing order of the previous season average. The figure shows that the two HB estimates are very close to each other and performed much better than the direct estimates. Table 3 reports the four summary statistics for both models and it further confirms the closeness of the two HB estimators.

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

Model	AAD	AARD
Bernoulli-Logit-EP	0.0195	0.0774
Bernoulli-Logit-Normal	0.0198	0.0790

The true values of p_{ins} are available for the baseball data. In order to investigate the nature of the random effects v_i, we fitted the logistic regression model defined by (3.1) on p_{ins} considering the previous season average at bats as the covariate. We then tested the normality of the residuals v_i using K-S normality test and the normal Q-Q plot. The

p-value of 0.676 from the K-S test concludes that v_i appear to be normal. One player was identified as outlier through the normal Q-Q plot. Excluding that outlier, v_i look approximately normal. The posterior mean of the kurtosis parameter φ estimated using the *Bernoulli-Logit-EP* model equals to 0.506. It further confirms the approximate normality of v_i .

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

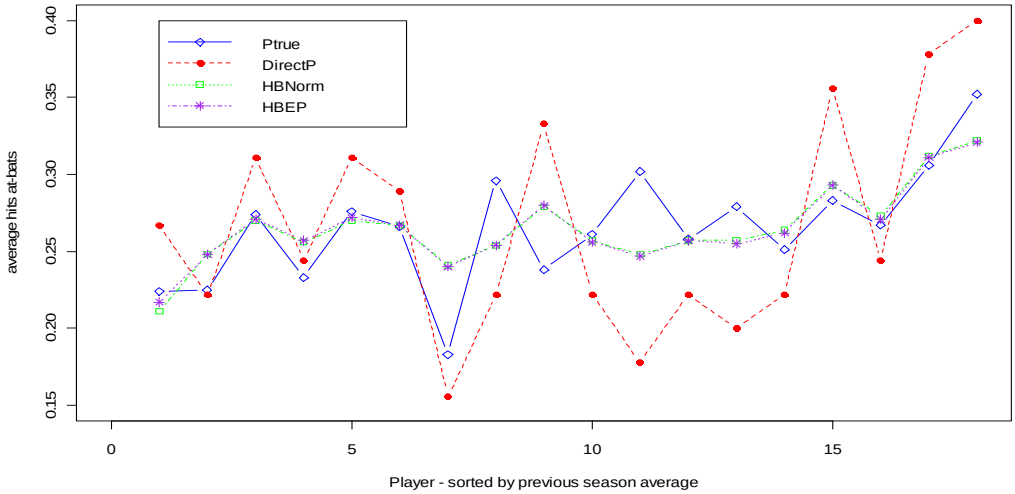

Figure 4: HB estimates of the batting average for the rest of the 1970 season

6.2.3 Missouri Turkey Hunting Survey Data

The Missouri Turkey Hunting Survey (MTHS) is a bi-annual postseason mail survey conducted by the Missouri Department of Conservation to monitor and aid in the regulation of the turkey hunting season. Questionnaires are mailed to a random sample of permit buyers after the turkey hunting season. The MTHS provides information concerning the number of turkeys harvested by hunters on each day of the hunting season and the total number of trips made to the counties by these hunters on each hunting day. The success rates are then obtained from this information. The 1994 spring season data has been analyzed by He and Sun (1998). The problem was to estimate the county level success rates P_i for all counties in Missouri. They provided hierarchical Bayesian estimates of success rates for all the 114 counties in Missouri using a simple Binomial-Beta model without covariates.

We revisit the 1994 spring season data analyzed by He and Sun (1998) in this subsection. There were three counties with zero sample size. They could be predicted from the same model using the parameters estimated from the rest data. We excluded them from our analysis for simplicity purpose. The sample sizes for the rest 111 counties varied from 2 to 802. We computed the HB estimates for the rest 111 counties using the two models *Bernoulli-Logit-EP* and *Bernoulli-Logit-Normal* with no covariates under the same prior assumptions considered in the earlier sections. Figure 5 displays the two HB estimates (HBEP and HBNorm) along with the direct sample estimates (DirectP) sorted by the sample size in the increasing order. Figure 6 displays the corresponding standard

errors of the point estimates. The two HB estimates appeared close to each other for many of the counties, with HBEP being a little bit closer to the direct estimates than the HBNorm. The graph clearly shows that when the sample size is small, the deviation between the direct estimates and the HB estimates is large. But as the sample size is getting larger, the deviation is getting smaller. For the county with the largest sample size ($n_i = 802$), all the three estimates become the same.

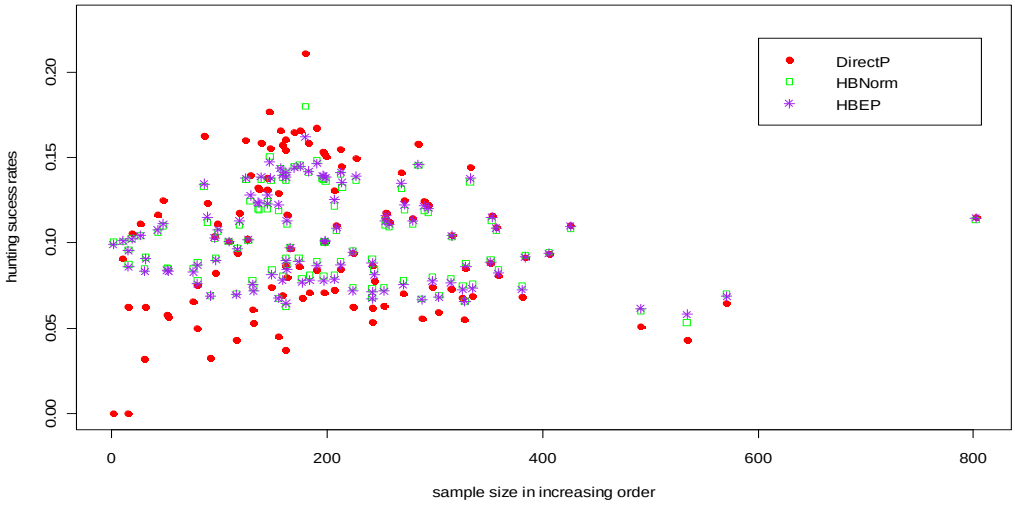

Figure 5: Estimation of the Turkey Hunting success rates

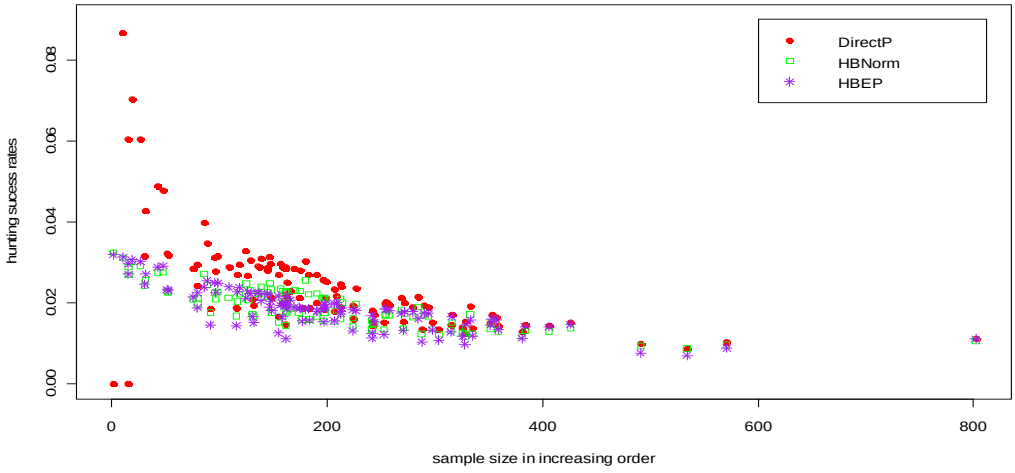

Figure 6: Standard errors of the direct estimates and posterior standard errors of the HB estimates of the Turkey hunting success rates

Figure 7 displays the posterior density plots for the hyperparameters σ and φ. The upper two panels present the standard deviation σ and the kurtosis φ from the *Bernoulli-Logit-EP* model respectively. The lower panel presents σ from the *Bernoulli-Logit-Normal* model. Both of the two plots on the left show bell shapes for σ, though the estimates from the EP model appearing a more sharp shape than the other one. The

posterior density plot for φ shows that the mode of φ is around 0.05. The posterior mean of the kurtosis parameter φ is around 0.2. All these evidences indicate that the random effects v_i are platikurtik and therefore *Bernoulli-Logit-EP* may be a more appropriate model to fit this data.

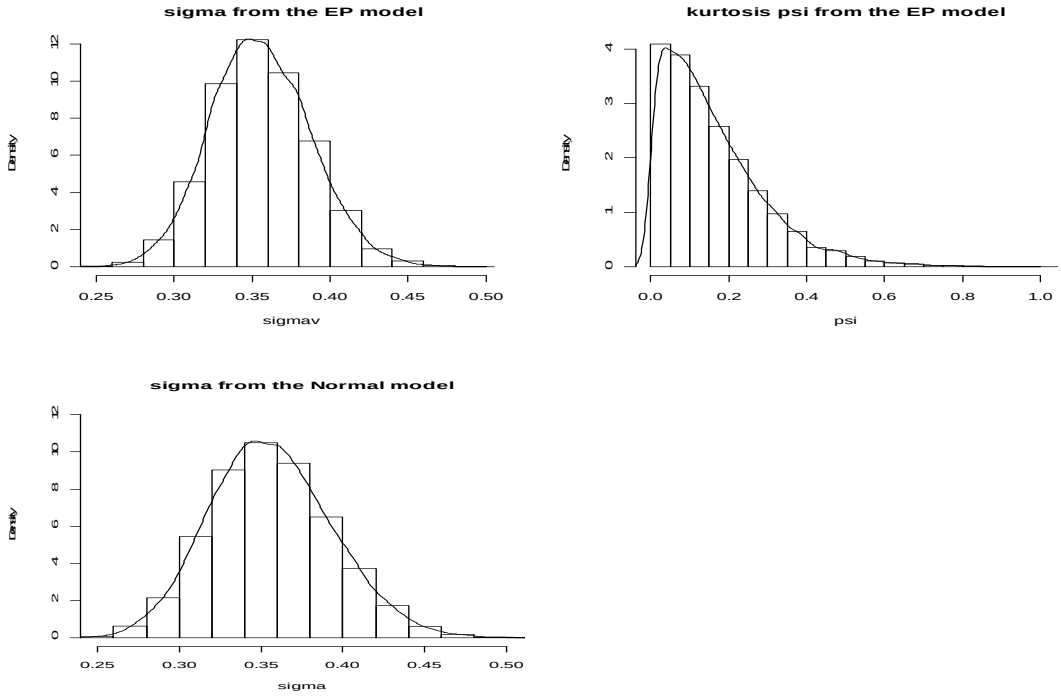

Figure 7: Posterior density of the hyperparameters

7. Concluding Remarks

The proposed *Bernoulli-Logit-EP* model extended the usual logistic regression mixed model by assuming a class of probability distributions in modeling the distribution of random effects. We considered an adaptive approach in which the shape parameter (φ) is automatically determined by the survey data. The parameter φ is 0.5 under normality. Our empirical data analyses based on both simulated data and real survey data demonstrated the robustness of the *Bernoulli-Logit-EP* model and suggested that the proposed model works efficiently to accommodate potential kurtosis and zero problems. To avoid computation burden, we only generated a few samples in our evaluation study based on simulated data. So the evaluation results are limited.

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

Acknowledgements

The author would like to thank her Ph.D. supervisory committee members Drs Partha Lahiri, Graham Kalton, Keith Rust, Paul Smith, and Richard Valliant for their generous guidance and help through her dissertation research.

References

Box, G.E.P. and Tiao, G.C. (1973). *Bayesian Inference in Statistical Analysis*. Addison-Wesley, Reading MS.

Chen, M., Shao, Qi. and Ibraham, J.G. (2000). *Monte Carlo methods in Bayesian computation.* New York: Springer-Verlag.

Choy, S.T.B. and Smith, A.F.M. (1997a). On robust analysis of a normal location parameter. Journal of Royal Statistical Society, Series B, 59, 463-474.

Choy, S.T.B. and Smith, A.F.M. (1997b). Hierarchical models with scale mixtures of normal distributions. TEST 6, 202-221.

Efron, B. and Morris, C.N. (1975). Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association*, 70, 311-319.

Fabrizi, E. and Trivisano, C. (2007). Robust models for mixed effects in linear mixed models applied to small area estimation. Submitted to *Journal of Statistical Planning and Inference.*

Farrell, P.J., MacGibbon, B., and Tomberlin, T.J. (1994). Protection against outliers in empirical Bayes estimation. *Canadian Journal of Statistics*, 22, 365-376.

Farrell, P.J., MacGibbon, B., and Tomberlin, T.J. (1997a). Empirical Bayes estimators of small area proportions in multistage designs. *Statistical Sinica*, 7, 1065-1083.

Farrell, P.J., MacGibbon, B., and Tomberlin, T.J. (1997b). Empirical Bayes small-area estimation using logistic regression models and summary statistics. *Journal of Business & Economic Statistics*, Vol. 15, 1, pp. 101-108.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1, 515-533.

Gelman, A., Carlin, J.B., Carlin, Stern, H.S., and Rubin, D.B. (1995). *Bayesian Data Analysis*, Chapman & Hall/CRC.

Ghosh, M. and Lahiri, P. (1987). Robust empirical Bayes estimation of means from stratified samples. *Journal of the American Statistical Association, Vol. 82, 1153-1162.*

He, Z. and Sun, D. (1998). Hiearchical Bayes estimation of hunting success rates. *Environmental and Ecological Statistics*, 5, 223-236.

Hogg, R.V. (1974). Adaptive robust procedures: A partial review and some suggestions for future applications and theory. *Journal of the American Statistical Association, 69, 909-923.*

Jiang, J., and Lahiri, P. (2006a). Mixed model prediction and small area estimation (with discussions). *Test*, 15, 1, 1-96.

Jiang, J., and Lahiri, P. (2006b). Estimation of finite population domain means: A model-assisted empirical best prediction approach. *Journal of the American Statistical Association*, 101, 301-311.

Li, Y, and Lahiri, P. (2007). Robust model-based and model-assisted predictors of the finite population total. *Journal of the American Statistical Association*, 102, 664-673.

Liu, B., Lahiri, P., and Kalton, G. (2008). Hierarchical Bayes modeling for survey-weighted small area proportions. Submitted to *Survey Methodology.*

MacGibbon, B., and Tomberlin, T.J. (1989), Small area estimates of proportions via Empirical Bayes Techniques. *Survey Methodology*, 15, 237-252.

Malec, D., Sedransank, J., Moriarity, C.L., and Lecler, F.B. (1997). Small area inference for binary variables in National Health Interview Survey. *Journal of the American Statistical Association*, 92, 815-826.

Prescott, P. (1978). Selection of trimming proportions for robust adaptive trimmed means. *Journal of the American Statistical Association, 73, 133-136.*

Rao, J.N.K. (2003). *Small area estimation*. New York: John Wiley and Sons.

Robert, C.P., and Casella, G. (1999). *Monte Carlo Statistical Methods*. New York: Springer-Verlag.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., and Van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society B*, 64: 583-640.