STATISTICS IN MEDICINE
Statist. Med. 2007; **26**:2699–2715
Published online 3 October 2006 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.2709

Estimation of the proportion of overweight individuals in small areas—a robust extension of the Fay–Herriot model

Dawei Xie[1,*,†], Trivellore E. Raghunathan[2] and James M. Lepkowski[2]

[1]*Department of Biostatistics and Epidemiology, Center for Clinical Epidemiology and Biostatistics, University of Pennsylvania School of Medicine, 617 Blockley Hall, 423 Guardian Drive, Philadelphia, PA 19104, U.S.A.*
[2]*Department of Biostatistics and Institute for Social Research, University of Michigan, U.S.A.*

SUMMARY

Hierarchical model such as Fay–Herriot (FH) model is often used in small area estimation. The method might perform well overall but is vulnerable to outliers. We propose a robust extension of the FH model by assuming the area random effects follow a t distribution with an unknown degrees-of-freedom parameter. The inferences are constructed using a Bayesian framework. Monte Carlo Markov Chain (MCMC) such as Gibbs sampling and Metropolis–Hastings acceptance and rejection algorithms are used to obtain the joint posterior distribution of model parameters. The procedure is used to estimate the county-level proportion of overweight individuals from the 2003 public-use Behavioral Risk Factor Surveillance System (BRFSS) data. We also discuss two approaches for identifying outliers in the context of this application. Copyright © 2006 John Wiley & Sons, Ltd.

KEY WORDS: t distribution; hierarchical model; complex sample survey; BRFSS; overweight; outlier detection

1. INTRODUCTION

For estimating the income for small places with population less than 1000, Fay and Herriot [1] generalized the James-Stein estimator to a regression model. In the Fay–Herriot (FH) model, both the design-based direct estimates of small area means and the area-level random effects are assumed to be normally distributed. Because of the central limit theorem, the distributional assumption on the design-based direct estimates, at least for moderately large samples, is easy to justify. Comparatively, the assumption on the random effects is hard to check and might be vulnerable to outliers. However, Lahiri and Rao [2] showed that a second-order approximation of mean square error (MSE) of the

*Correspondence to: Dawei Xie, Department of Biostatistics and Epidemiology, Center for Clinical Epidemiology and Biostatistics, University of Pennsylvania School of Medicine, 617 Blockley Hall, 423 Guardian Drive, Philadelphia, PA 19104, U.S.A.
†E-mail: dxie@cceb.med.upenn.edu

Received 31 August 2005
Accepted 31 July 2006

empirical best linear unbiased prediction (EBLUP) of the FH estimates in Prasad and Rao [3] is robust with respect to non-normality of the random effects under some regularity conditions. For example, they showed that Prasad-Rao MSE is robust when the distribution of the random effects is a t with at least 9 degrees-of-freedom.

The FH model has been extended to consider other distributions for random effects. For example, Datta and Lahiri [4] introduced a model assuming the random effects follow a scale mixture of normal distributions where the t distribution is a special case. Assuming the parameters of the scale mixture of normal distribution are known, they theoretically derived hierarchical Bayes small area estimates. This article extends the Datta and Lahiri approach by treating the degrees-of-freedom as an unknown parameter. This methodology is applied to obtain county-level estimates of the proportion of overweight individuals.

Section 2 discusses the proposed model and inference. Section 3 describes the details on model fitting in a Bayesian framework. Section 4 describes the results from a practical application with Behavioral Risk Factor Surveillance System (BRFSS) data [5]. Section 5 concludes with a discussion of advantages and limitations of the proposed model and future research plans.

2. EXTENSION TO FAY–HERRIOT MODEL AND INFERENCE

Let y_i and $d_i, i=1,\ldots,n$ be the design-based direct estimate of the population quantity and associated variance estimate accounting for the complex design for area i. The sampling distribution of y_i is assumed as

$$y_i \sim \mathrm{N}(\theta_i, d_i) \tag{1}$$

where θ_i is the quantity of our interest, the true population mean for area i.

In the FH model, it is further assumed that

$$\theta_i \sim \mathrm{N}(\mathbf{x}_i \boldsymbol{\beta}, \sigma^2) \tag{2}$$

where $\boldsymbol{\beta}$ is a vector of regression coefficients associated with area-level covariates $\mathbf{x}_i$ and σ^2 is the variance of the area-level random effects (or between-area variance). We propose to replace (2) with

$$\theta_i \sim t_v(\mathbf{x}_i \boldsymbol{\beta}, \sigma^2) \tag{3}$$

where $p(\theta) \sim t_v(\mu, \sigma^2)$ denotes a t distribution with location parameter μ, scale parameter σ^2, degrees-of-freedom v with the density

$$p(\theta|\mu, \sigma^2, v) = \frac{\Gamma\left(\dfrac{v+1}{2}\right)}{\Gamma\left(\dfrac{v}{2}\right)\sqrt{\pi v \sigma^2}} \left(1 + \frac{(\theta-\mu)^2}{v\sigma^2}\right)^{-(v+1)/2}$$

The t distribution is a member of a family of scale mixtures of normal distributions, where $\theta_i \sim \mathrm{N}(\mathbf{x}_i \boldsymbol{\beta}, u_i)$, $u_i = \sigma^2/\chi^2_v$, and χ^2_v is a χ^2 random variable with v degrees-of-freedom. Another popular member of the normal scale mixture is when u_i follows a Bernoulli distribution. In this case, the distribution of θ_i is a mixture of two normal distributions. The t distribution has been used in other settings. For example, Lange *et al.* [6] discussed the use of the t distribution for

DOI: 10.1002/sim

error terms in linear and non-linear regressions. Pinheiro *et al.* [7] discussed the application of multivariate t distributions in linear mixed effects models assuming the same degrees-of-freedom for the t distributions of the error term and random effects.

The likelihood of the model (1), (3) is therefore

$$L = \prod_{i=1}^{n} p(y_i|\theta_i, \boldsymbol{\beta}, \sigma^2, v) = \prod_{i=1}^{n} [p(y_i|\theta_i) p(\theta_i|\boldsymbol{\beta}, \sigma^2, v)]$$

$$= \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi d_i}} \exp\left(-\frac{(y_i-\theta_i)^2}{2d_i}\right) \frac{\Gamma\left(\frac{v+1}{2}\right)}{\Gamma\left(\frac{v}{2}\right)\sqrt{\pi v \sigma^2}} \left(1 + \frac{(\theta_i - \mathbf{x}_i\boldsymbol{\beta})^2}{v\sigma^2}\right)^{-(v+1)/2} \right]$$

The likelihood given above is analytically intractable. As we will show later, the formality of the t distribution with random scale mixture will be computationally very useful when u_i are treated as missing data, as data augmentation can then be used to draw values from the relevant posterior distribution.

2.1. Estimation of θ_i when σ^2, $\boldsymbol{\beta}$, and v are known

When σ^2, $\boldsymbol{\beta}$, and v are known, the posterior mean of θ_i is given by

$$\hat{\theta}_i = y_i + \frac{E^{h_i}[h_i f(h_i)]}{E^{h_i}[f(h_i)]} = w_i y_i + (1 - w_i)\mathbf{x}_i\boldsymbol{\beta} \tag{4}$$

where

$$h_i \sim \mathrm{N}(0, d_i), \quad f(h_i) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\Gamma\left(\frac{v}{2}\right)\sqrt{\pi v \sigma^2}} \left(1 + \frac{(y_i + h_i - \mathbf{x}_i\boldsymbol{\beta})^2}{v\sigma^2}\right)^{-(v+1)/2}$$

$$w_i = E_p^{u_i}\left(\frac{1/d_i}{\frac{1}{d_i} + \frac{1}{u_i}}\right)$$

and $E_p^{u_i}$ is with respect to the marginal posterior density of u_i:

$$p(u_i|y_i, \boldsymbol{\beta}, \sigma^2, v) \propto p(y_i|\boldsymbol{\beta}, \sigma^2, v, u_i) p(u_i|v, \sigma^2)$$

$$= \frac{1}{\sqrt{2\pi(d_i + u_i)}} \exp\left(-\frac{(y_i - \mathbf{x}_i\boldsymbol{\beta})^2}{2(d_i + u_i)}\right) p(u_i|v, \sigma^2)$$

where $p(u_i|v, \sigma^2)$ is the density of $\mathrm{Inv} - \chi^2(v, \sigma^2)$ which is

$$p(u_i|v, \sigma^2) = \frac{(v/2)^{v/2}}{\Gamma(v/2)} \sigma^v u_i^{-(v/2+1)} \exp\left(\frac{-v\sigma^2}{2u_i}\right)$$

DOI: 10.1002/sim

2.2. *Estimation of θ_i when σ^2, $\boldsymbol{\beta}$, and v are unknown*

When σ^2, $\boldsymbol{\beta}$, and v are unknown, it is possible to estimate these parameters via maximum likelihood (ML) or restricted maximum likelihood (REML) approaches. Then an empirical Bayes (EB) estimator of θ_i can be obtained by replacing σ^2, $\boldsymbol{\beta}$, and v in (4) with the ML or REML estimates. However, (4) does not have a closed-form solution, and numerical integration has to be applied. Computing the MSE of $\hat{\theta}_i$ is difficult. Furthermore, conditional on the estimates $\hat{\sigma}^2$, $\hat{\boldsymbol{\beta}}$, and $\hat{v}$ as if they are true values may underestimate the MSE of $\hat{\theta}_i$. The improved MSE estimate could be obtained by using resampling techniques such as the jackknife or bootstrap [8–10].

3. ESTIMATION OF θ_i USING FULLY BAYESIAN APPROACH

We have adopted a fully Bayesian approach to base inferences on the marginal posterior distribution $p(\theta_i|\text{data})$. For the fully Bayesian approach we need prior distribution for hyperparameters σ^2, $\boldsymbol{\beta}$, and v. We assume independent priors for the three hyperparameters, i.e. $p(\boldsymbol{\beta}, \sigma^2, v) = p(\boldsymbol{\beta})p(\sigma^2)p(v)$. Following Gelman *et al.* [11], we assume an improper uniform prior for σ^2 and $\boldsymbol{\beta}$. Following Watanabe [12], a proper prior for v is assumed with $p(\boldsymbol{\beta}, \sigma^2, v) \propto p(v)$ where $v \sim \text{Gamma}(\alpha, \gamma)$, and $\text{Gamma}(\alpha, \gamma)$ denotes a Gamma distribution with mean α/γ and variance α/γ^2 for known α and γ.

The marginal posterior distributions of σ^2, $\boldsymbol{\beta}$, v, or θ_i cannot be written explicitly. However, the joint posterior distribution can be simulated using a Markov Chain Monte Carlo (MCMC) such as Gibbs sampling [13, 14] or the Metropolis–Hastings algorithm [11]. Note that when v is assumed known, we could extend Raghunathan and Rubin [15] by generating the sequence under a normal distributional assumption using Gibbs sampling and then using the importance sampling to redraw values from the generated sequence.

By treating u_i as missing data, the implementation of MCMC becomes easier. The conditional distributions of σ^2, $\boldsymbol{\beta}$, or θ_i involve normal, inverse χ^2, or Gamma distributions. The conditional distribution of v is not standard. We adapt a Metropolis–Hastings acceptance–rejection algorithm proposed by Watanabe [12]. For a general discussion on this algorithm, see Chib and Greenberg [16]. The conditional distributions involved in the MCMC steps are given in the appendix.

4. COUNTY-SPECIFIC PROPORTION OF OVERWEIGHT INDIVIDUALS FROM THE BEHAVIORAL RISK FACTOR SURVEILLANCE SYSTEM

We illustrate the model inference by estimating the county-level proportions of overweight individuals using the 2003 public-use BRFSS data [5]. We also discuss two approaches for detecting outliers and compare the model estimates from the FH and t models.

4.1. *Data*

The data come from two separate sources. One is the public-use data of the BRFSS, a telephone survey on the health behaviours of US adults, conducted by the Centers for Disease Control and Prevention (CDC). In this data example, we obtain the direct estimate y_i, the proportion of overweight individuals, and associated design-based variance estimate d_i, on the county level in the United States. Operationally, an adult is said to be 'overweight' if his or her body mass index

DOI: 10.1002/sim

(BMI) is over 25 where the BMI is defined as weight in kilograms divided by the square of height in meters.

In the public-use BRFSS data, only 1051 counties, the state of Alaska, and the District of Columbia can be identified. The identifiers for the rest of the 2061 counties were suppressed due to confidentiality concerns. The 1053 'counties' (treating Alaska and the District of Columbia as akin to counties) comprise 81.1 per cent of the total population in the nation with a total sample size of 200 810. The y_i's range from 0.308 to 0.819 with a median of 0.604, while standard errors range from 0.011 to 0.143 with a median of 0.055.

The second data source is the 2000 Census from which we choose four county-level covariates ($\mathbf{x}_i$). These four covariates are per cent of Hispanic population (x_{1i}), per cent of people who have a bachelor or higher degree among those 25 years or over (x_{2i}), percentage of individuals taking public transportation to work for workers 16 years and over (x_{3i}), and the percentage of population that is 0–18 years old (x_{4i}). We also considered many other variables related to the county's urban/rural status, Metropolitan Statistical Area (MSA) status, and various characteristics of the population such as employment status, medium income, poverty level, and per cent of blue-collar workers. They did not contribute much when adjusting for the other four covariates included in the model. The four covariates are included on the log-scale to reduce the impact of skewness, and they are standardized to reduce the burden in computing $(\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i')^{-1}$.

4.2. *Implementation of MCMC*

To obtain the Bayesian estimates under the FH and t models, programs were written in the GAUSS programming language to run 2000 iterations for each of 10 independent sequences of Gibbs sampler. After a burn-in period of the first 1000 iterations in each sequence, the convergence of the Gibbs sampler was assessed using Gelman–Rubin statistic R [17] for every component in θ_i, σ^2, $\boldsymbol{\beta}$ and v. The initial values were selected from a series of uniform distributions to ensure dispersion. For example, the initial values of θ_i, σ^2, $\boldsymbol{\beta}$ and v are from uniform distributions on $[0, 1]$, $(0, 1]$, $[-20, 20]$, and $(0, 20]$, respectively. The range of θ_i is easy to determine since it should be between 0 and 1. For the degrees-of-freedom v, it is almost equivalent to a normal model when $v>20$. Therefore, a uniform distribution between 0 and 20 for v is good enough to cover all of the desired possibilities. We also tried other uniform distributions with much wider ranges for these parameters, and it did not substantially change the data analysis results.

We chose 10 independent sequences following Gelman and Rubin [17]. The rate of convergence for the normal model is much higher than that of the corresponding t model. For the normal model, all R's are smaller than 1.1 after 20 iterations. For the corresponding t model, it takes 360 iterations for all R's to be smaller than 1.1.

4.3. *Identifying outlying counties under normal model*

We first fit the FH model to the data. The posterior mean of θ_i ranges from 0.435 to 0.701 with a mean of 0.608, while the posterior variance of θ_i ranges from 0.0001 to 0.0010.

Under the FH (normal) model, the marginal distribution of y_i by integrating θ_i out is $y_i \sim \mathrm{N}(\mathbf{x}_i\boldsymbol{\beta}, \sigma^2+d_i)$. As suggested in Dempster and Ryan [18], we have $\delta_i(\sigma^2, \boldsymbol{\beta}) \equiv (y_i - \mathbf{x}_i\boldsymbol{\beta})/\sqrt{\sigma^2+d_i} \sim \mathrm{N}(0, 1)$ where $\delta_i^2(\sigma^2, \boldsymbol{\beta})$ is a Mahalanobis-like distance, as defined in Lange *et al.* [6]. When σ^2 and $\boldsymbol{\beta}$ are replaced by the maximum likelihood estimate (MLE) $\tilde{\sigma}^2$ and $\tilde{\boldsymbol{\beta}}$, asymptotically $\delta_i(\tilde{\sigma}^2, \tilde{\boldsymbol{\beta}}) \sim \mathrm{N}(0, 1)$ too. In a Bayesian setting with a weakly informative or non-informative prior, the posterior mode is approximately equivalent to the MLE. For the FH model, the posterior mode can further be

DOI: 10.1002/sim

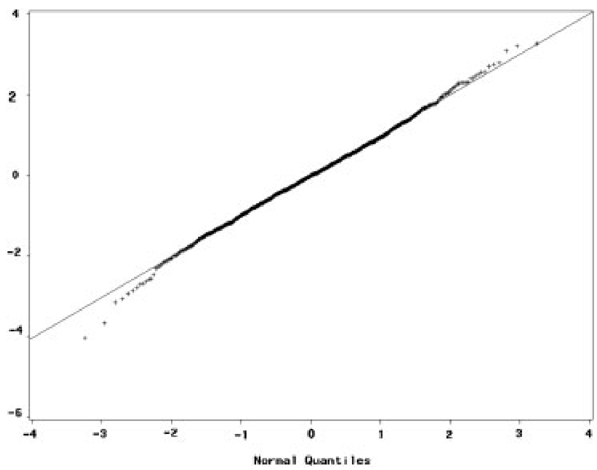

Figure 1. QQ plot of $\delta_i(\hat{\sigma}^2, \hat{\boldsymbol{\beta}})$ under FH (normal) model. The straight line is the expected line when there are no outliers.

approximated by the posterior means $\hat{\sigma}^2$ and $\hat{\boldsymbol{\beta}}$, and thus $\delta_i(\hat{\sigma}^2, \hat{\boldsymbol{\beta}})$ approximately follows a standard normal distribution. Therefore, $\delta_i(\hat{\sigma}^2, \hat{\boldsymbol{\beta}})$ can be used to check the assumptions in the FH model. Note that $\delta_i(\hat{\sigma}^2, \hat{\boldsymbol{\beta}})$ considers not only the distance between the direct estimate y_i and the regression synthetic estimate $\mathbf{x}_i\hat{\boldsymbol{\beta}}$, but also the reliability of the direct estimate y_i, d_i. Only those areas with 'extreme' and relatively reliable direct estimates (i.e. large $|y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}|$ and small d_i) will be recognized as outliers.

There is another way to identify outliers. As in You and Rao [19] and Daniels and Gatsonis [20], we can simulate the posterior predictive distribution of a hypothetical replication of the direct estimates. Computationally, drawing from the posterior predictive distribution is nearly effortless given that we have the draws of θ_i from its posterior distribution. For every draw of θ_i, we simulate a hypothetical replicate direct estimate from $y_i^{\text{rep}} \sim \text{N}(\theta_i, d_i)$. The resulting draws of y_i^{rep} represent the posterior predictive distribution of y_i.

Define a p-value for each county under the FH model as

$$
\begin{aligned}
p_i^{\text{N}} &= \frac{1}{J}\sum_{j=1}^{J} I[y_i^{\text{rep}(j)} > y_i] \\
I[y_i^{\text{rep}(j)} > y_i] &= \begin{cases} 1 & \text{if } y_i^{\text{rep}(j)} > y_i \\ 0 & \text{otherwise} \end{cases}
\end{aligned}
\tag{5}
$$

where $j = 1, \ldots, J$ indexes the number of replicates and J is the number of draws of y_i^{rep}. For county i, a p_i^{N} close to 0.5 indicates a good fit of the model, a value close to 0 or 1 indicates a lack of fit of the model. A county with $p_i^{\text{N}} < 0.05$ or $p_i^{\text{N}} > 0.95$ might be viewed as an outlier.

Figure 1 shows the QQ plot of $\delta_i(\hat{\sigma}^2, \hat{\boldsymbol{\beta}})$ under the FH (normal) model with each point representing a county. The two counties at the lower left corner show a violation of the normal distribution

DOI: 10.1002/sim

Table I. Outlying counties and their estimates (the numbers in the parentheses are associated standard error or posterior standard derivations).

County	y_i	$\hat{\theta}_i$ from normal	$\delta_i(\hat{\sigma}^2, \hat{\boldsymbol{\beta}})$	p_i^{N}	$\hat{\theta}_i$ from t, $v=3.96$	p_i^{t}
A	0.361 (0.076)	0.618 (0.030)	−3.674	0.952	0.549 (0.091)	0.856
B	0.309 (0.059)	0.522 (0.029)	−4.044	0.962	0.412 (0.078)	0.774

County A: Logan County, KT; County B: Park County, MT.

Table II. Posterior means and standard derivations (in parentheses) of σ^2 and $\boldsymbol{\beta}$ from the normal model and t model with $v=3.96$.

	County level covariates	Normal model	t model
$\boldsymbol{\beta}$	Hispanic population (%)	−0.0065 (0.0022)	−0.0070 (0.0025)
	Taking public transportation to work in workers 16 years of age or older (%)	−0.0060 (0.0022)	−0.0060 (0.0022)
	Population 0–18 years old (%)	0.0144 (0.0019)	0.0148 (0.0021)
	Bachelor degree or higher education in those 25 years of age or older (%)	−0.0278 (0.0018)	−0.0276 (0.0019)
σ^2		0.0010 (0.0001)	0.0005 (0.0001)

assumption. Both outlying counties have a $\delta_i(\hat{\sigma}^2, \hat{\boldsymbol{\beta}})$ smaller than −3.5, and p-values p_i^{N}'s greater than 0.95. Table I gives the direct estimate, the posterior mean from the normal model, the posterior mean from a t model, $\delta_i(\hat{\sigma}^2, \hat{\boldsymbol{\beta}})$, and p_i^{N} for each of the two outlying counties, namely Logan County, Kentucky and Park County, Montana. The two counties have direct estimates 0.3–0.4 with a small standard error. However, no county-level covariates can explain the low direct estimates in these two counties. The four county-level covariates x_{1i}, x_{2i}, x_{3i}, and x_{4i} (see Table II for the explanations on the covariates) are 1.0, 0.3, 25.7 and 9.6 per cent for Logan County, Kentucky, 2.0, 1.4, 23.5 and 23.1 per cent for Park County, Montana, compared to national estimates 12.5, 4.8, 27.1 and 24.3 per cent. Therefore, the proportion of overweight individuals is predicted to be higher than the national level in Logan County, Kentucky, and close to the national level in Park County, Montana.

4.4. The t model

A t model with unknown degrees-of-freedom v is then applied to the data as one approach to accommodating outliers. The parameters $\alpha=\gamma=10^{-4}$ were chosen in the Gamma prior for v so that it is only weakly informative. The posterior distribution of v is shown in Figure 2. The posterior mean of v is 3.96 with a standard deviation 0.16 with a 95 per cent posterior confidence interval (3.62, 4.24). Although it appears a little right-skewed, the mode of v is very close to the posterior median 4.00.

DOI: 10.1002/sim

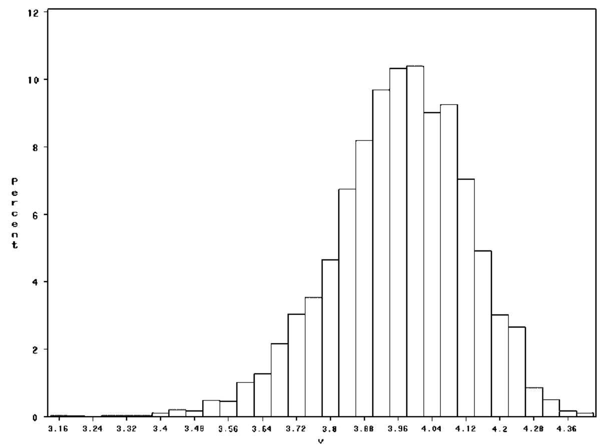

Figure 2. Histogram of draws of degree-of-freedom v from its posterior distribution after burn-in period in Gibbs sampling.

4.5. Comparison of estimates from the FH and t models

4.5.1. Small area estimates. We compare the posterior distribution of θ_i and one striking difference is in the skewness. In the FH model the posterior distribution of θ_i is approximately symmetric, while in the t model it is skewed except for those with a p_i^{N} close to 0.5. Figure 3 shows histograms from selected counties (including the two outlying counties) with p_i^{N} from 0.300 to 0.962. We observe that the distribution of θ_i is left skewed when $p_i^{\mathrm{N}}<0.5$ and right skewed when $p_i^{\mathrm{N}}>0.5$.

Figure 4(a) shows the posterior mean of θ_i from the FH model and the t model. For the majority of the counties, the differences are small. The shrinkage effect is smaller in the t model, and $\hat{\theta}_i$ agrees well when the direct estimate y_i is in the middle of the marginal distribution of θ_i values (around 0.6). The differences between the two model estimates are mostly between -0.05 and 0.03 except for the two outlying counties. In addition, we can observe that the differences are a monotone function of $\delta_i(\hat{\sigma}^2, \hat{\boldsymbol{\beta}})$, as shown in Figure 4(b).

The posterior standard derivations of θ_i are between 0.010 and 0.143 with a median 0.054 for the t model, while those for the normal model are between 0.010 and 0.031 with a median 0.028. Table II shows that the posterior standard deviations for the two outlying counties are larger under the t model than under the normal. One explanation of the high variation is because these draws are mostly from one tail of the t distribution.

4.5.2. Estimates of σ^2 and $\boldsymbol{\beta}$. Table II gives the estimates of σ^2 and $\boldsymbol{\beta}$. Note that the interpretations of σ^2 are different in the normal and t models. Therefore, it is not meaningful to compare σ^2 under different models.

The posterior mean and standard deviation of $\boldsymbol{\beta}$ from the two models do not differ much. From both models, we can see that percentage of bachelor degrees or higher education in the population aged 25 years or over ($\mathbf{x}_4$) and percentage of population 0–18 years old ($\mathbf{x}_3$) explain more variation

DOI: 10.1002/sim

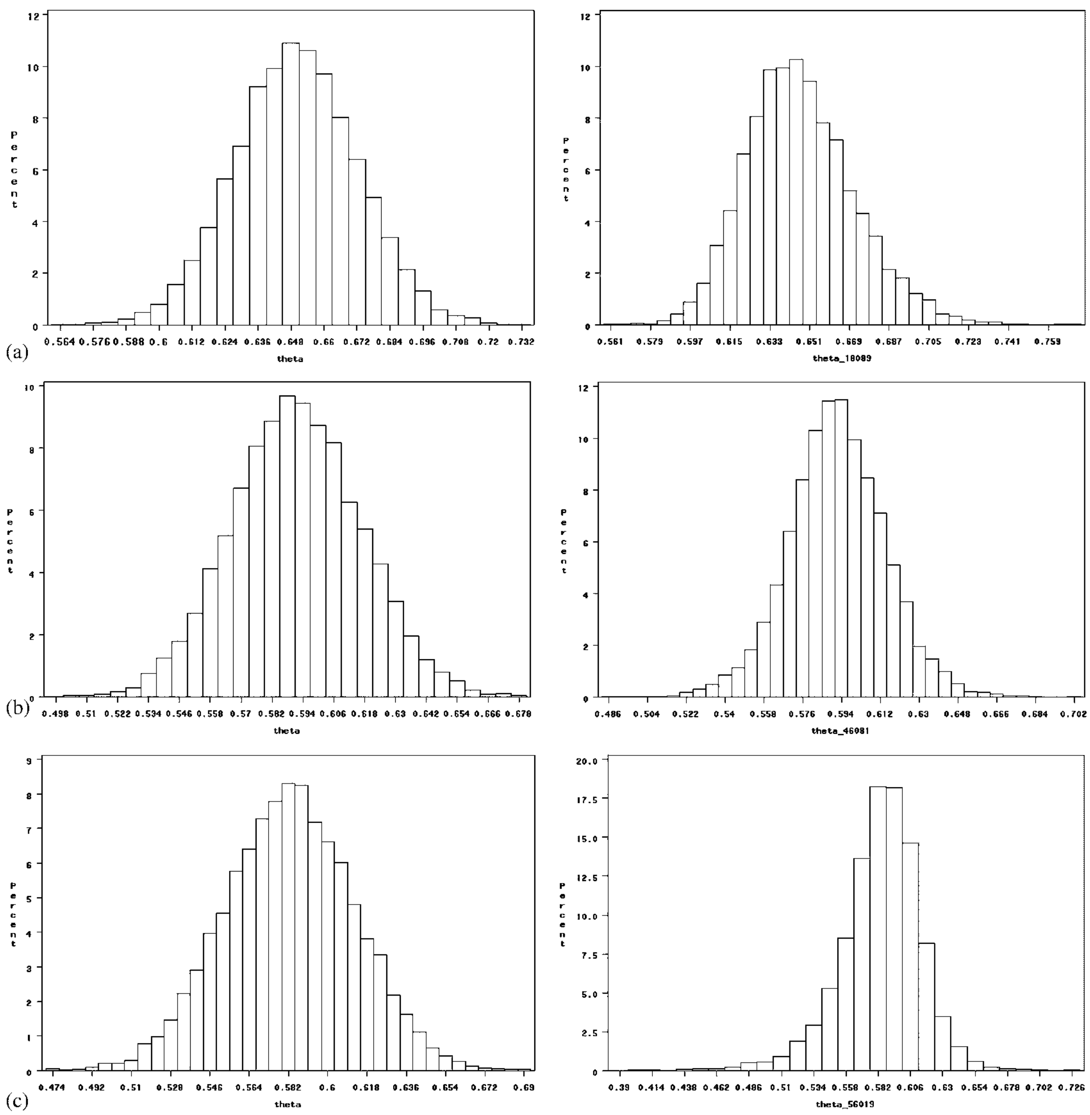

Figure 3. Posterior distribution of θ_i from normal and t models for selected counties (the histogram on the left is from the normal model, on the right is from the t model): (a) posterior distribution of θ_i from the normal model and t model for a county with $p_i^{\mathrm{N}} = 0.3$; (b) posterior distribution of θ_i from the normal model and t model for a county with $p_i^{\mathrm{N}} = 0.5$; (c) posterior distribution of θ_i from the normal model and t model for a county with $p_i^{\mathrm{N}} = 0.7$; and (d) posterior distribution of θ_i from the normal model and t model for a county with $p_i^{\mathrm{N}} = 0.952$; and (e) posterior distribution of θ_i from the normal model and t model for a county with $p_i^{\mathrm{N}} = 0.962$.

DOI: 10.1002/sim

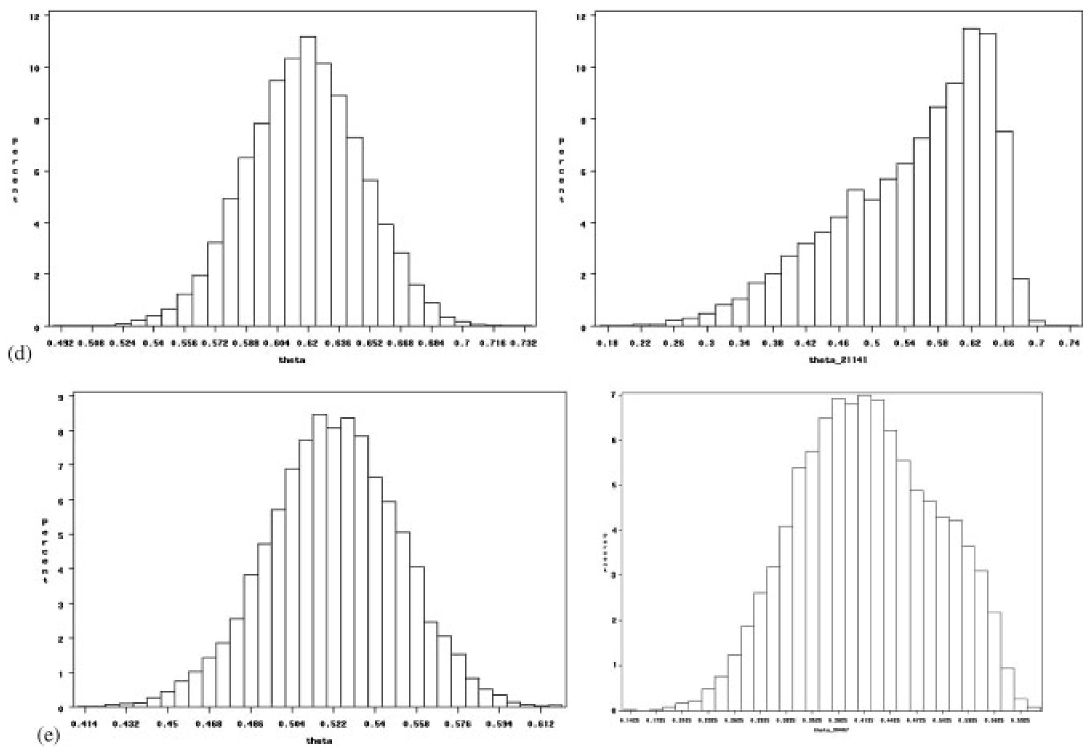

Figure 3. *Continued*

in the county-level proportion of overweight individuals than the other two county covariates. The proportion of overweight individuals is positively correlated with the percentage of population 0–18 years old ($\mathbf{x}_3$), and negatively correlated with percentage of bachelor degrees or higher education in people 25 years old or over ($\mathbf{x}_4$), per cent of Hispanic population ($\mathbf{x}_1$), and per cent of taking public transportation to work among workers 16 years old or over ($\mathbf{x}_2$).

4.6. Goodness of fit of the FH and t models

In Section 10.2.6, Rao [21] discussed MCMC methods such as Bayes factors, posterior predictive densities, and cross-validation predictive densities for model selection in a Bayesian framework. Interested readers can further refer to Gelfand [22] and Berger and Pericchi [23] for details.

We used two approaches to check the model fit. As in Datta *et al*. [24], we calculated the divergence measure proposed by Laud and Ibrahim [25]:

$$d(\hat{\boldsymbol{\theta}}, \mathbf{y}) = \frac{1}{nK}\sum_{k}^{K} \|\hat{\boldsymbol{\theta}}^{(k)} - \mathbf{y}\|^2 = \frac{1}{nK}\sum_{k}^{K}\sum_{i}^{n} (\hat{\theta}_i^{(k)} - y_i)^2$$

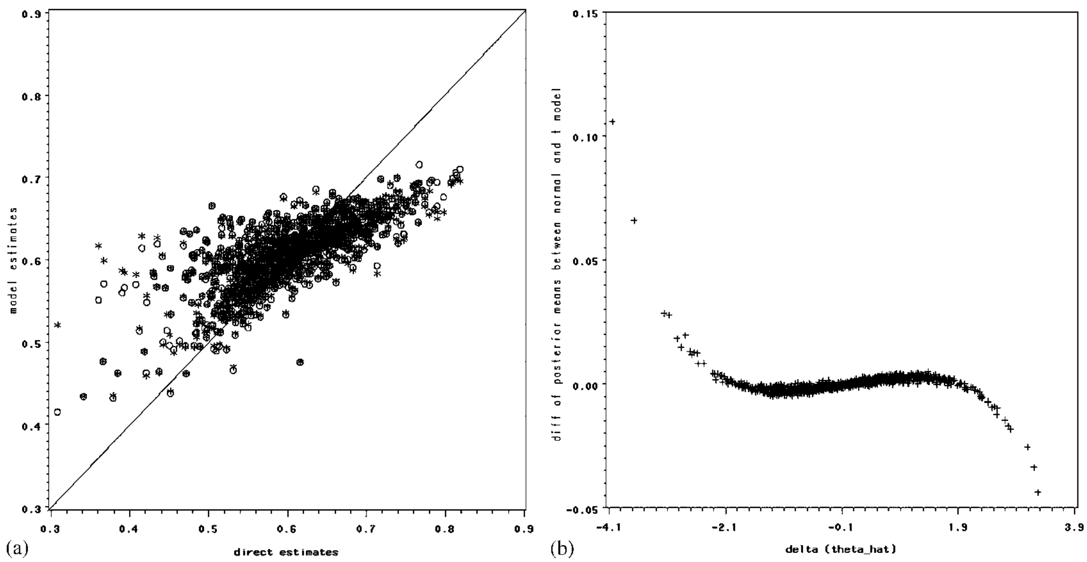

Figure 4. (a) Posterior mean $\hat{\theta}_i$ from the normal, and t model *versus* direct estimates. The line is what to expect if there is no shrinkage effect. The stars are posterior mean from the normal model, the open circles are from the t model; and (b) differences between the posterior mean $\hat{\theta}_i$ from the normal and t model and $\delta_i(\hat{\sigma}^2, \hat{\boldsymbol{\beta}})$. The two points on the left upper corner correspond to the two outlying counties.

where $\hat{\theta}_i^{(k)}$ is the kth draw ($k = 1, \ldots, K$) of θ_i in the MCMC. The divergence measure $d(\hat{\boldsymbol{\theta}}, \mathbf{y})$ can be decomposed into two components, $d(\hat{\boldsymbol{\theta}}, \mathbf{y}) = (1/nK)\sum_k^K \sum_i^n [(\hat{\theta}_i^{(k)} - \hat{\theta}_i)^2 + (\hat{\theta}_i - y_i)^2]$ where $\hat{\theta}_i$ is the posterior mean of θ_i. The first component $(1/nK)\sum_k^K \sum_i^n (\hat{\theta}_i^{(k)} - \hat{\theta}_i)^2$, denoted as $d_1(\hat{\boldsymbol{\theta}}, \mathbf{y})$, is the posterior variance of θ_i and the second $(1/nK)\sum_k^K \sum_i^n (\hat{\theta}_i - y_i)^2 = (1/n)\sum_i^n (\hat{\theta}_i - y_i)^2$, denoted as $d_2(\hat{\boldsymbol{\theta}}, \mathbf{y})$, is the distance between the posterior mean and the direct estimate y_i. In general, the divergence measure is similar in the FH model ($d(\hat{\boldsymbol{\theta}}, \mathbf{y}) = 0.0034$) and the t model ($d(\hat{\boldsymbol{\theta}}, \mathbf{y}) = 0.0033$). The t model has a slightly larger $d_1(\hat{\boldsymbol{\theta}}, \mathbf{y})$ and smaller $d_2(\hat{\boldsymbol{\theta}}, \mathbf{y})$ ($d_1(\hat{\boldsymbol{\theta}}, \mathbf{y})$=0.00071, $d_2(\hat{\boldsymbol{\theta}}, \mathbf{y})$=0.0027) than the FH model ($d_1(\hat{\boldsymbol{\theta}}, \mathbf{y}) = 0.00070$, $d_2(\hat{\boldsymbol{\theta}}, \mathbf{y}) = 0.0026$). This is consistent with Figures 4(a) and (b) where it is shown that most of the posterior means of θ_i do not differ much between the two models.

Although both models fit the overall data well, the two p_i^{N}'s greater than 0.95 indicate that the FH model does not fit some of the counties well. To check the model fit for the t model, we can define the p-value for each county under the t model as in (5). Denoting this p-value p_i^{t} we find that p_i^{t}'s are between 0.166 and 0.879 while p_i^{N}'s for the normal model are between 0.088 and 0.962.

4.7. Sensitivity analysis

It is important to check the sensitivity of the data analysis with respect to the choice of prior for the hyperparameters. Changing the values of α and γ in the Gamma prior to 10^{-3} or 10^{-5} does not substantially change the data analysis results. We can also change α to 10^{-3} and γ to

2×10^{-4}, 10^{-4}, or 5×10^{-5} so that the prior mean for v equal to 5, 10 or 20. These choices of hyperparameters do not substantially change the analysis results either. The maximum absolute relative change is 0.5 per cent in posterior mean and 10 per cent in posterior standard deviation of θ_i. The changes in the posterior mean and standard deviation of other hyperparameters are less than 2 per cent.

Another choice is a proper prior for σ^2 and $\boldsymbol{\beta}$, such as $p(\boldsymbol{\beta}) \sim \mathrm{N}(0, \boldsymbol{\Sigma}_0)$,

$$p(\sigma^2) \sim \mathrm{Inv}-\mathrm{Gamma}(a_0, b_0) \equiv \mathrm{Inv}-\chi^2\left(2a_0, \frac{b_0}{a_0}\right)$$

i.e.

$$p(\sigma^2) = \frac{b_0^{a_0}}{\Gamma(a_0)}(\sigma^2)^{-(a_0+1)}\mathrm{e}^{-b_0/\sigma^2}$$

The hyperparameter $\boldsymbol{\Sigma}_0$ can be chosen as $c_0\mathbf{I}_p$ where $\mathbf{I}_p$ is a $p\times p$ identity matrix where p is the total number of elements in $\boldsymbol{\beta}$. In order to have a proper prior for σ^2, we need $a_0 \geqslant 1$ in the inverse χ^2 prior. For the prior to be diffuse, c_0 can be a very large positive number, e.g. $c_0 = 10^4$, a_0 can be a small number, e.g. $a_0 = 1$, and b_0 can be a very small positive number, e.g. $b_0 = 10^{-4}$.

Note that for the prior of v, when $\alpha = 1$, it is the exponential prior proposed in Geweke [26] and Fernandez and Steel [27]. As noted in Watanabe [12], this prior is restrictive since the mean and variance are $1/\gamma$ and $1/\gamma^2$, respectively.

In Gelman *et al.* [11], an uniform prior is assumed for $1/v$ on [0,1], which is equivalent to an improper prior on v, i.e. $p(v) \propto 1/v^2$, $v > 1$. It is equivalent to the gamma prior when $\alpha = -1$ and $\gamma = 0$. Although it is not a special case of Gamma prior since α and γ are positive for a Gamma distribution, we can follow the same argument in Watanabe [12] to conclude that the conditional distribution of v is unimodal when $n > 4$.

In general, we should consider refining the choice of prior if the inference depends strongly on some particular parameters, for example, σ^2 or v. Under this situation, one can first obtain an empirical Bayesian estimator of θ_i treating $\boldsymbol{\beta}$, σ^2 and v fixed to explore which parameters have the most impact on estimating θ_i. See Gelman *et al.* [11] for a general discussion.

We performed sensitivity analysis by constructing inferences about θ_i under various prior distributions. The differences were quite minimal suggesting that the choice of diffuse proper or improper prior has little effect on the estimates.

5. DISCUSSION

In this article, we extended the FH model by using a t distribution to model the small area mean where there are outliers. Compared to the FH model, there are three properties of the t model with unknown degrees-of-freedom that are arguably advantageous:

1. There is less shrinkage in the t model. When σ^2 and $\boldsymbol{\beta}$ are known, the posterior mean of θ_i from the FH model is given by $\hat{\theta}_i^{\mathrm{FH}} = w_i^{\mathrm{FH}} y_i + (1 - w_i^{\mathrm{FH}})\mathbf{x}_i\boldsymbol{\beta}$ where $w_i^{\mathrm{FH}} = 1/d_i/(1/d_i + 1/\sigma^2) = \sigma^2/(\sigma^2 + d_i)$. The estimate $\hat{\theta}_i^{\mathrm{FH}}$ is called the 'shrinkage' estimator because it shrinks the direct estimator y_i toward the synthetic estimator $\mathbf{x}_i\boldsymbol{\beta}$. The degree of shrinkage can also be measured by w_i^{FH} in the sense that a high w_i^{FH} corresponds to a lower

degree of shrinkage. Note that the weight w_i^{FH} in the FH model does not depend on the direct estimator y_i or the regression-synthetic estimator $\mathbf{x}_i\boldsymbol{\beta}$. Therefore, as long as two direct estimators have equal d_i's, the degree of shrinkage in $\hat{\theta}_i^{\mathrm{FH}}$ will be the same. However, for a direct estimator with larger $|y_i - \mathbf{x}_i\boldsymbol{\beta}|$, we will expect a larger $|y_i - \hat{\theta}_i^{\mathrm{FH}}|$, which is the distance between the direct estimator and the shrinkage estimator $\hat{\theta}_i^{\mathrm{FH}}$. In many cases, this property is not ideal. In the t model the shrinkage to the extreme direct estimators is limited. From (4), we can see that $\hat{\theta}_i$ from the t model can also be expressed as a weighted average of y_i and $\mathbf{x}_i\boldsymbol{\beta}$. Further analysis can show that w_i, the weight associated with the direct estimator y_i in the t model, is a non-decreasing function of $|y_i - \mathbf{x}_i\boldsymbol{\beta}|$ given σ^2 and d_i. This suggests that when two direct estimators have the same sampling variance, the estimator $\hat{\theta}_i$ from the t model gives more weight to the direct estimator with larger $|y_i - \mathbf{x}_i\boldsymbol{\beta}|$ while a large $|y_i - \mathbf{x}_i\boldsymbol{\beta}|$ occurs when $\mathbf{x}_i\boldsymbol{\beta}$ does not predict y_i well.

When σ^2 and $\boldsymbol{\beta}$ are unknown, $\hat{\theta}_i$ is not a convex combination of y_i and $\mathbf{x}_i\boldsymbol{\beta}$ any more since w_i is a function of y_i and $\mathbf{x}_i\boldsymbol{\beta}$ for any finite v. However, the data example shows that the weight on the direct estimate y_i is on average higher in a t model than that in the normal model.

2. The degree of shrinkage is decided by data, as we assumed the degrees-of-freedom v unknown. We observe that w_i is a non-increasing function of v. This indicates the weight on a given direct estimate y_i with in a t model is higher than that in a normal model when given σ^2 and $\boldsymbol{\beta}$. The subjectivity of fixing the degrees-of-freedom in advance is avoided.
3. The posterior standard deviations of small area estimates incorporate the uncertainty of the degrees-of-freedom naturally in the fully Bayesian framework.

There are also some limitations of the t model approach and some areas that seem to require further study.

1. Our model assumes independent sampling errors among areas. Due to the complex nature of the sampling designs, the sampling errors are sometimes correlated, as when the small areas cut across clusters in the sampling design. Under this situation, one might consider a multivariate model for y_i, i.e. $\mathbf{y} \sim \mathrm{N}(\boldsymbol{\theta}, \mathbf{D})$ where $\mathbf{y} = (y_1, \ldots, y_n)'$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)'$, $\mathbf{D}$ is the sampling variance matrix of $\mathbf{y}$. This is the case discussed in Datta and Lahiri [4]. Under a Bayesian framework the extension is theoretically straightforward. With the model for $\boldsymbol{\theta}$ unchanged, the conditional distributions of all parameters are unchanged except for $\boldsymbol{\theta}$. However, the computation might be intensive.
2. We assume a normal model for the direct estimate y_i although y_i is the mean of binary variables indicating whether a sampled person is overweight. The normal assumption might not hold when the sample size is small. When ignoring the complex sampling design, one can assume a binary model for $n_i y_i$, i.e. $n_i y_i \sim \mathrm{Binomial}(n_i, \theta_i)$, where n_i is the sample size from area i. We can further assume $\mathrm{logit}(\theta_i) \sim t_v(\mathbf{x}_i\boldsymbol{\beta}, \sigma^2)$. Rao [21] discussed the normal case where $\mathrm{logit}(\theta_i) \sim \mathrm{N}(\mathbf{x}_i\boldsymbol{\beta}, \sigma^2)$ in Section 10.11.2. The conditional distribution of θ_i is not standard and algorithms such as Metropolis–Hastings can be used to obtain the joint posterior distributions. When the survey sampling design is complex, further investigation is needed to investigate how to take it into consideration. A possibility is to use the effective sample size in each area to replace the actual sample size in $n_i y_i \sim \mathrm{Bin}(n_i, \theta_i)$.

DOI: 10.1002/sim

APPENDIX

Under the t model, the conditional distributions of the model parameters are:

(1)
$$\boldsymbol{\beta}|y_i, \mathbf{x}_i, \theta_i, \sigma^2, u_i, i=1,\ldots,n \sim \mathrm{N}\left(\tilde{\boldsymbol{\beta}}, \left(\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i'}{u_i}\right)^{-1}\right)$$

where

$$\tilde{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n}\frac{\theta_i\mathbf{x}_i'}{u_i}\right)\left(\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i'}{u_i}\right)^{-1}$$

(2)
$$\sigma^2|y_i, \mathbf{x}_i, \theta_i, \boldsymbol{\beta}, u_i,\ i=1,\ldots,n \sim \mathrm{Gamma}\left(\frac{nv+1}{2}, v\sum_{i=1}^{n}\frac{1}{2u_i}\right)$$

(3)
$$\theta_i|y_i, \mathbf{x}_i, \boldsymbol{\beta}, \sigma^2, u_i,\ i=1,\ldots,n \sim \mathrm{N}\left(\frac{\dfrac{y_i}{d_i}+\dfrac{\mathbf{x}_i\boldsymbol{\beta}}{u_i}}{\dfrac{1}{d_i}+\dfrac{1}{u_i}}, \frac{1}{\dfrac{1}{d_i}+\dfrac{1}{u_i}}\right)$$

if y_i is not missing

$$\theta_i|y_i, \mathbf{x}_i, \boldsymbol{\beta}, \sigma^2, u_i,\ i=1,\ldots,n \sim \mathrm{N}(\mathbf{x}_i\boldsymbol{\beta}, u_i)$$

if y_i is missing

(4)
$$u_i|y_i, \mathbf{x}_i, \theta_i, \boldsymbol{\beta}, \sigma^2,\ i=1,\ldots,n \sim \mathrm{In}\, v-\chi^2\left(v+1, \frac{1}{v+1}[(\theta_i-\mathbf{x}_i\boldsymbol{\beta})^2+v\sigma^2]\right)$$

(5) Conditional distribution of v:

$$v|y_i, \mathbf{x}_i, \boldsymbol{\beta}, \sigma^2, u_i,\ i=1,\ldots,n \propto p(v)\prod_{i=1}^{n}p(u_i|v, \sigma^2)$$

where

$$p(v)=\frac{\gamma^{\alpha}}{\Gamma(\alpha)}v^{\alpha-1}\exp(-\gamma v) \quad \text{and} \quad u_i|v, \sigma^2 \sim \mathrm{In}\, v-\chi^2(v, \sigma^2)$$

i.e.

$$p(u_i|v, \sigma^2)=\frac{(v/2)^{v/2}}{\Gamma(v/2)}(\sigma^2)^{v/2}u_i^{-(v/2+1)}\exp\left(-\frac{v\sigma^2}{2u_i}\right)$$

Therefore,

$$\ln f(v|y_i, \mathbf{x}_i, \boldsymbol{\beta}, \sigma^2, u_i,\ i=1,\ldots,n) = \text{const.}+\frac{nv}{2}\ln\left(\frac{v}{2}\right)-n\Gamma\left(\frac{v}{2}\right)-\eta v+(\alpha-1)\ln v$$

where

$$\eta=\frac{1}{2}\sum_{i=1}^{n}\left(\ln\frac{u_i}{\sigma^2}+\frac{\sigma^2}{u_i}\right)+\gamma$$

DOI: 10.1002/sim

Watanabe [12] proved that this conditional distribution is unimodal if $2\alpha + n > 2$, which is satisfied as long as the number of areas $n > 2$.

Suppose there is a candidate-generating distribution $h(v)$ such that it is possible to sample directly from $h(v)$ by some known method. Watanabe [12] proposed to use a normal distribution as $h(v)$ with mean $v^* - A/B$ and variance $-1/B$ where

$$A = \left.\frac{\partial \ln f(v)}{dv}\right|_{v=v^*} = \frac{n}{2}\left\{\ln\frac{v^*}{2} + 1 - \psi\left(\frac{v^*}{2}\right)\right\} - \eta + \frac{\alpha - 1}{v^*}$$

$$B = \left.\frac{\partial^2 \ln f(v)}{dv^2}\right|_{v=v^*} = \frac{n}{2}\left\{\frac{1}{v^*} - \frac{1}{2}\psi'\left(\frac{v^*}{2}\right)\right\} - \frac{\alpha - 1}{v^{*2}}$$

$\psi(v) = \partial \ln \Gamma(v)/\partial v$, and $\psi'(v) = \partial\psi(v)/\partial v$. Note that $\psi(v)$ and $\psi'(v)$ are called psi (digamma) and trigamma function, respectively.

Let $\ln f^*(v) = (nv/2)\ln(v/2) - n\Gamma(v/2) - \eta v + (\alpha - 1)\ln v$, and $\ln h^*(v) = (nv^*/2)\ln(v^*/2) - n\Gamma(v^*/2) - \eta v^* + (\alpha - 1)\ln v^* + A(v - v^*) + (B/2)(v - v^*)^2$. Denote the jth sampled value of v by v_j and consider the $(j + 1)$th sampling. The Metropolis–Hastings algorithm is as follows.

a. Sample a candidate v_x from the candidate-generating distribution $h(v)$and a value r_1 from the uniform distribution on (0, 1).
b. If $r_1 \leqslant f^*(v_x)/h^*(v_x)$, return v_x; else, go to a.
c. If $f^*(v_j) < h^*(v_j)$, then let $q = 1$;
If $f^*(v_j) \geqslant h^*(v_j)$ and $f^*(v_x) < h^*(v_x)$, then let $q = h^*(v_j)/f^*(v_j)$;
If $f^*(v_j) \geqslant h^*(v_j)$ and $f^*(v_x) \geqslant h^*(v_x)$, then let $q = \min\{[f^*(v_x)h^*(v_j)]/[f^*(v_j)h^*(v_x)], 1\}$;
d. Sample a value r from the uniform distribution on (0, 1).
e. If $r \leqslant q$, return $v_{j+1} = v_x$. Else, return $v_{j+1} = v_j$.

To speed up the algorithm, the value of v^* is selected to solve

$$A = \left.\frac{\partial \ln f(v)}{dv}\right|_{v=v^*} = \frac{n}{2}\left\{\ln\frac{v^*}{2} + 1 - \psi\left(\frac{v^*}{2}\right)\right\} - \eta + \frac{\alpha - 1}{v^*} = 0$$

The equation might be solved by standard methods. For example, starting from v_0^*, the Newton–Raphson algorithm involves setting the next $v_1^* = v_0^* - (A/B)|_{v^*=v_0^*}$, so on and so forth.

Under the FH model with a normal assumption on θ_i, the conditional distributions of the model parameters are

$$\boldsymbol{\beta}|y_i, \mathbf{x}_i, \theta_i, \sigma^2, \quad i = 1, \ldots, n \sim \mathrm{N}\left(\hat{\boldsymbol{\beta}}, \left(\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i'\right)^{-1}\sigma^2\right) \tag{6}$$

where

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} \theta_i\mathbf{x}_i'\right)\left(\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i'\right)^{-1}$$

$$\sigma^2|y_i, \mathbf{x}_i, \theta_i, \boldsymbol{\beta}, \quad i = 1, \ldots, n \sim \mathrm{In}\, v - \chi^2\left(n - 1, \frac{n}{n-1}s^2\right) \tag{7}$$

where

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(\theta_i - \mathbf{x}_i\boldsymbol{\beta})^2$$

$$\theta_i|y_i,\mathbf{x}_i,\boldsymbol{\beta},\sigma^2, \quad i=1,\ldots,n \sim \mathrm{N}\left(\frac{\dfrac{y_i}{d_i}+\dfrac{\mathbf{x}_i\boldsymbol{\beta}}{\sigma^2}}{\dfrac{1}{d_i}+\dfrac{1}{\sigma^2}}, \frac{1}{\dfrac{1}{d_i}+\dfrac{1}{\sigma^2}}\right) \tag{8}$$

if Y_i is not missing

$$\theta_i|y_i,\mathbf{x}_i,\boldsymbol{\beta},\sigma^2, \quad i=1,\ldots,n \sim \mathrm{N}(\mathbf{x}_i\boldsymbol{\beta},\sigma^2)$$

if Y_i is missing.

REFERENCES

1. Fay RE, Herriot RA. Estimates of income for small places: an application of James–Stein procedures to census data. *Journal of the American Statistical Association* 1979; **74**(366):269–277.
2. Lahiri P, Rao JNK. Robust estimation of mean squared error of small area estimators. *Journal of the American Statistical Association* 1995; **90**(430):758–766.
3. Prasad NGN, Rao JNK. The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association* 1990; **85**(409):163–171.
4. Datta GS, Lahiri P. Robust hierarchical Bayes estimation of small area characteristics in the presence of covariates and outliers. *Journal of Multivariate Analysis* 1995; **54**(2):310–328.
5. Behavioral Risk Factor Surveillance System. At http://www.cdc.gov/brfss with the last access on 7/24/06.
6. Lange KL, Little RJA, Taylor JMG. Robust statistical modeling using the t-distribution. *Journal of the American Statistical Association* 1987; **84**(403):881–896.
7. Pinheiro JC, Liu C, Wu YN. Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t distribution. *Journal of Computational and Graphical Statistics* 2001; **10**(2):249–276.
8. Jiang J, Lahiri P, Wan S. Jackknifing the mean squared error of empirical best predictor. *Annals of Statistics* 2002; **30**(6):1782–1810.
9. Arora V, Lahiri P, Mukherjee K. Empirical Bayes estimation of finite population means from complex surveys. *Journal of the American Statistical Association* 1997; **92**(440):1555–1562.
10. Laird NM, Louis TA. Empirical Bayes confidence intervals based on bootstrap samples. *Journal of the American Statistical Association* 1987; **82**(399):739–750.
11. Gelman A, Carlin JB, Stern HS, Rubin DB. *Bayesian Data Analysis*. Chapman & Hall: London, New York, 1995.
12. Watanabe T. On sampling the degree-of-freedom of Student's-t disturbances. *Statistics and Probability Letters* 2001; **52**(2):177–181.
13. Gelfand AE, Smith AFM. Gibbs sampling for marginal posterior expectations. *Communications in Statistics*, *Part A—Theory and Methods* 1991; **20**(5–6):1747–1766.
14. Tierney L. Exploring posterior distributions using Markov chains. In *Computing Science and Statistics*: *Proceedings of the 23rd Symposium on the Interface*, Keramidas EM (ed.). Interface Foundation: Fairfax Station, 1991; 563–570.
15. Raghunathan TE, Rubin DB. An application of Bayesian statistics using sampling/importance resampling for a deceptively simple problem in quality control. In *Data Quality Control*: *Theory and Pragmatics*, Liepins G, Uppuluri VRR (eds). Marcel Dekker: New York, 1990; 229–243.
16. Chib S, Greenberg E. Understanding the Metropolis–Hastings algorithm. *The American Statistician* 1995; **49**(4):327–335.
17. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. *Statistical Science* 1992; **7**(4):457–472.
18. Dempster AP, Ryan LM. Weighted normal plots. *Journal of the American Statistical Association* 1985; **80**(412): 845–850.
19. You Y, Rao JNK. Small area estimation using unmatched sampling and linking models. *Canadian Journal of Statistics* 2002; **30**(1):3–15.

DOI: 10.1002/sim

20. Daniels MJ, Gatsonis C. Hierarchical generalized linear models in the analysis of variation in heath care utilization. *Journal of the American Statistical Association* 1999; **94**(445):29–42.
21. Rao JNK. *Small Area Estimation*. Wiley: New York, 2003.
22. Gelfand AE. Models determination using sampling-based methods. In *Markov Chain Monte Carlo in Practice*, Gilks WR, Richarson S, Spiegelhalter DJ (eds). Chapman & Hall: London, 1996; 145–162.
23. Berger J, Pericchi L. Objective Bayesian methods for model selection: introduction and comparison. In *Model Selection*, Lahiri P (ed.). Institute of Mathematical Statistics: Beachwood Ohio, 2001; 135–207.
24. Datta GS, Lahiri P, Maiti T, Lu KL. Hierarchical Bayes estimation of unemployment rates for the states of the U.S. *Journal of the American Statistical Association* 1999; **94**(448):1074–1082.
25. Laud P, Ibrahim J. Predictive model selection. *Journal of the Royal Statistical Society*, *Series B* 1995; **57**(1):247–262.
26. Geweke J. Bayesian treatment of the independent Student-*t* linear model. *Journal of Applied Econometrics* 1993; **8**:19–40.
27. Fernandez C, Steel MFJ. On Bayesian modeling of fat tails and skewness. *Journal of the American Statistical Association* 1998; **93**(441):359–371.

DOI: 10.1002/sim