Estimation of Prevalence of Overweight in Small Areas -A Robust Extension of Fay-Herriot Model

Dawei Xie[1], Trivellore E. Reghunathan[2], James M. Lepkowski[2]
[1]Department of Biostatistics and Epidemiology, University of Pennsylvania
[2]Department of Biostatistics, University of Michigan

ABSTRACT

Hierarchical model such as Fay-Herriot (FH) model is often used to develop small area estimates. It 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 degree of freedom. The inference is done in a Bayesian framework. Monte Carlo Markov Chain (MCMC) techniques 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 illustrated in an example, in which we estimate the county level prevalence of overweight from the 2003 public-use Behavioral Risk Factor Surveillance System (BRFSS) data. We also applied two approaches to identify outliers in this application.

Key Words: *t* distribution; Hierarchical model; Complex sample survey; BRFSS; Overweight; Outlier detection.

1. Introduction

In the context of estimating the income for small places with population less than 1,000, Fay and Herriot (1979) generalized the James-Stein estimator to a regression model and obtained small area estimates. 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. The distribution assumption on the design-based direct estimates is easier to justify because of central limit theorem. Comparatively, the assumption on the random effects is hard to check and might be vulnerable to outliers. In this article we extend FH model by assuming a *t* distribution for the random effects or "between-area" effects. We call this model "*t* model" contrast to the "FH model" or "normal model" when random effects are assumed normally distributed.

The *t* distribution is important in robust statistical modeling. Lange, Little, and Taylor (1989) discussed the use of the *t* distribution for error terms in linear and nonlinear regressions. Pinheiro, Liu, and Wu (2001) discussed the application of multivariate *t* distribution in linear mixed effects models assuming the same degree of freedom for the *t* distributions of the error term and random effects.

In the literature of small area estimation, there are a number of related papers on the robust extensions of the FH model, and some used a class of models for the random effects where the *t* distribution is a special case. Lahiri and Rao (1995) showed that an estimator of mean square error (MSE) of the empirical best linear unbiased prediction (EBLUP) of the FH estimates in Prasad and Rao (1990) is robust with respect to nonnormality of the random effects under some regularity condition. Specifically, for *t* distribution the regularity condition requires that the degree of freedom is greater than 9. Datta and Lahiri (1995) introduced a model assuming the random effects follow a scale mixture of normal distribution where *t* distribution is a special case. Assuming the parameters of the scale-mixture normal distribution known, they derived the hierarchical Bayes small area estimates in theory. In this article we assumed the parameter of the *t* distribution, the degree of freedom, is unknown and obtained its posterior distribution. Further we applied this approach on a real data example.

This article is organized as follows. In section 2 we discuss the proposed model and inference. Section 3 describes the details in applying the extension to an example, including how to detect outliers and how to evaluate model fit. Discussion and future study are given Section 4.

2. Extention to Fay-Herriot Model and Inference

Let $y_i, i = 1, \cdots, n$ be the design-based direct estimates of the proportion mean of our interest in area *i*. Let d_i be the design-based variance estimate for y_i by taking into account the complex design features in

the survey. The sampling distribution of y_i is assumed as

$$y_i \sim N(\theta_i, d_i) \quad (1)$$

where θ_i is the quantity of our interest, the true population mean for area *i*. In FH model, it is further assumed that

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

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

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

where $t_v(\mu, \sigma^2)$ denotes a *t* distribution with location parameter μ, scale parameter σ^2, degree of freedom *v*, and density

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

Note that the t distribution is symmetric and the degree of freedom *v* can be any positive number. Specifically, the *t* distribution with 1 degree of freedom is known as the Cauchy distribution. The *t* distribution approaches the normal distribution as $v \to \infty$. The *t* distribution belongs to a family of scale mixture of normal distributions as discussed in Datta and Lahiri (1995). This family of models can be represented by $\theta_i \sim N(\mathbf{X}_i\boldsymbol{\beta}, u_i)$ and u_i follows a distribution with longer-than-normal tails. Specifically, when u_i follows a Bernoulli distribution, θ_i is a mixture of two normal distributions. θ_i follows a *t* distribution $\theta_i \sim t_v(\mathbf{x}_i\boldsymbol{\beta}, \sigma^2)$ when $u_i \sim \text{Inv}-\chi^2(v, \sigma^2)$ where $\text{Inv}-\chi^2(v, \sigma^2)$ denotes a scaled inverse Chi-square distribution with mean $\frac{v}{v-2}\sigma^2$ and variance $\frac{2v^2}{(v-2)^2(v-4)}\sigma^2$.

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

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

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

2.1 When σ^2, $\boldsymbol{\beta}$, and *v* are known

When σ^2, $\boldsymbol{\beta}$, and *v* are known, i.e., (3) is a prior for θ_i, it is easy to show that the posterior mean of θ_i, as in Datta and Lahiri (1995), is

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

where $h_i \sim N(0, d_i)$,

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

$w_i = E_p^{u_i}\left(\frac{4n_i}{4n_i+\frac{1}{u_i}}\right)$, and $E_p^{u_i}$ is with respect to the marginal posterior density of u_i:

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

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

where $p(u_i \mid v, \sigma^2)$ is the density of $\text{Inv}-\chi^2(v, \sigma^2)$.

Note that from (4), $\hat{\theta}_i$ is a convex combination of y_i and $\mathbf{x}_i\boldsymbol{\beta}$, i.e., 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 estimate y_i, is a nondecreasing function of $|y_i - \mathbf{x}_i\boldsymbol{\beta}|$ and a nonincreasing function of v. This indicates the weight on direct estimate y_i in a *t* model is higher than that in a normal model when given σ^2 and $\boldsymbol{\beta}$. When σ^2 and $\boldsymbol{\beta}$ are unknown, $\hat{\theta}_i$ is not a convex of y_i and $\mathbf{x}_i\boldsymbol{\beta}$ any more since w_i is a function of y_i and $\mathbf{x}_i\boldsymbol{\beta}$ when $v < +\infty$. The weight on the direct estimate y_i is not necessarily higher in a *t* model for some *i*.

2.2 When σ^2, $\boldsymbol{\beta}$, and *v* are unknown

When σ^2, $\boldsymbol{\beta}$, and *v* are unknown, one way is to estimate these parameters is 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 and numerical integration has to be applied. The computation of the variance is not readily available either.

We propose to obtain the joint posterior distribution of all the unknown parameters, $\sigma^2, \boldsymbol{\beta}$, *v*, and θ_i, under a fully Bayesian framework. Following Gelman et al (1995), we assume an improper uniform prior for σ^2 and $\boldsymbol{\beta}$, and a vague proper prior for *v*, i.e., $p(\boldsymbol{\beta}, \sigma^2, v) \propto p(v) = Gamma(\alpha, \gamma)$ where $Gamma(\alpha, \gamma)$ denotes a Gamma distribution with mean α/γ and variance α/γ^2.

The joint posterior distribution of unknown parameters is then

$$
\begin{aligned}
&p(\theta_i, \boldsymbol{\beta}, \sigma^2, v, i = 1, \cdots, n \mid y_i, d_i, \mathbf{x}_i) \\
&\propto p(\boldsymbol{\beta}, \sigma^2, v) \prod_{i=1}^{n} p(y_i \mid \theta_i, \boldsymbol{\beta}, \sigma^2, v) \\
&= p(v) \prod_{i=1}^{k} \left[p(y_i \mid \theta_i)\, p(\theta_i \mid \boldsymbol{\beta}, \sigma^2, v) \right] p(v) \\
&= p(v) \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi d_i}} \exp\left(-\frac{(y_i - \theta_i)^2}{2 d_i} \right) \right. \\
&\qquad \left. \frac{\Gamma(\frac{v+1}{2})}{\Gamma(\frac{v}{2})\sqrt{\pi v \sigma^2}} \left(1 + \frac{(\theta_i - \mathbf{x}_i \boldsymbol{\beta})^2}{v \sigma^2} \right)^{-\frac{v+1}{2}} \right]
\end{aligned}
$$

The marginal posterior distribution of $\sigma^2, \boldsymbol{\beta}$, *v*, or θ_i cannot be written explicitly. However, the joint posterior distribution can be simulated using the Markov Chain Monte Carlo technique such as Gibbs sampling (Gelfand and Smith (1991), Tierney (1991)) and Metropolis-Hastings algorithm. Note that when *v* is assumed known, Raghunathan and Rubin (1990) provided an importance resampling algorithm to obtain the joint posterior distribution.

The conditional distributions of $\sigma^2, \boldsymbol{\beta}$, or θ_i involve normal, inverse Chi-square, or Gamma distributions. The conditional distribution of *v* is not standard. We adapt a Metropolis-Hastings acceptance-rejection algorithm proposed by Watanabe (2001). For a general discussion on this algorithm, see Chib and Greenberg (1995). The conditional distributions are as following.

(5) $\boldsymbol{\beta} \mid y_i, \mathbf{x}_i, \theta_i, \sigma^2, u_i, i = 1, \cdots, n \sim N\left(\tilde{\boldsymbol{\beta}}, \left(\sum_{i=1}^{k} \frac{\mathbf{x}_i \mathbf{x}_i'}{u_i} \right)^{-1} \right)$ where $\tilde{\boldsymbol{\beta}} = \left(\sum_{i=1}^{k} \frac{\theta_i \mathbf{x}_i'}{u_i} \right) \left(\sum_{i=1}^{k} \frac{\mathbf{x}_i \mathbf{x}_i'}{u_i} \right)^{-1}$;

(6) $\sigma^2 \mid y_i, \mathbf{x}_i, \theta_i, \boldsymbol{\beta}, u_i, i = 1, \cdots, n \sim Gamma\left(\frac{kv+1}{2}, v \sum_{i=1}^{n} \frac{1}{2u_i} \right)$;

(7) $\theta_i \mid y_i, \mathbf{x}_i, \boldsymbol{\beta}, \sigma^2, u_i, i = 1, \cdots, n \sim N\left(\frac{4n_i y_i + \frac{\mathbf{x}_i \boldsymbol{\beta}}{u_i}}{4n_i + \frac{1}{u_i}}, \frac{1}{4n_i + \frac{1}{u_i}} \right)$

if y_i is not missing;

$\theta_i \mid y_i, \mathbf{x}_i, \boldsymbol{\beta}, \sigma^2, u_i, i = 1, \cdots, n \sim N(\mathbf{x}_i \boldsymbol{\beta}, u_i)$ if y_i is missing;

(8) $u_i \mid y_i, \mathbf{x}_i, \theta_i, \boldsymbol{\beta}, \sigma^2, i = 1, \cdots, n \sim Inv\text{-}\chi^2\left(v+1, \frac{1}{v+1}\left[(\theta_i - \mathbf{x}_i \boldsymbol{\beta})^2 + v\sigma^2 \right] \right)$.

(9) Conditional distribution of *v*:

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

where $p(v) = \frac{\gamma^\alpha}{\Gamma(\alpha)} v^{\alpha-1} \exp(-\gamma v)$,

and $u_i \mid v, \sigma^2 \sim \text{Inv} - \chi^2(v, \sigma^2)$, i.e.,

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

Therefore,

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

where $\eta = \frac{1}{2} \sum_{i=1}^{n} \left(\ln \frac{u_i}{\sigma^2} + \frac{\sigma^2}{u_i} \right) + \gamma$. Watanabe (1999) 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 (1999) proposed to use a normal distribution as $h(v)$ with mean $v^* - A/B$ and variance $-1/B$ where

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

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

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

Let

$$\ln f^*(v)=\frac{nv}{2}\ln\left(\frac{v}{2}\right)-n\Gamma\left(\frac{v}{2}\right)-\eta v+(\alpha-1)\ln v,$$ and

$$\ln h^*(v)=\frac{nv^*}{2}\ln\left(\frac{v^*}{2}\right)-n\Gamma\left(\frac{v^*}{2}\right)-\eta v^*+(\alpha-1)\ln v^* + A(v-v^*)+\frac{B}{2}(v-v^*)^2 .$$

Denote the *j*th sampled value of v by v_j and consider the $(j+1)^{th}$ sampling. The Metropolis-Hastings algorithm is:

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 $r_1 \le \frac{f^*(v_x)}{h^*(v_x)}$, return v_x; else, go to a.

c. If $f^*(v_j) < h^*(v_j)$, then let *q*=1;
If $f^*(v_j) \ge h^*(v_j)$
and $f^*(v_x) < h^*(v_x)$, then let
$q=\frac{h^*(v_j)}{f^*(v_j)}$;
If $f^*(v_j) \ge h^*(v_j)$
and $f^*(v_x) \ge h^*(v_x)$, then let
$q=\min\left\{\frac{f^*(v_x)h^*(v_j)}{f^*(v_j)h^*(v_x)},1\right\}$;

d. Sample a value *r* from the uniform distribution on (0, 1).

e. If $r \le q$, 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 set the next $v^*_1=v^*_0-\left.\frac{A}{B}\right|_{v^*=v^*_0}$, so on and so forth.

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

(10) $$\boldsymbol{\beta} \mid y_i,\mathbf{x}_i,\theta_i,\sigma^2,i=1,\cdots,n \sim N\left(\hat{\boldsymbol{\beta}},\left(\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\sigma^2\right)$$

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

(11) $$\sigma^2 \mid y_i,\mathbf{x}_i,\theta_i,\boldsymbol{\beta},i=1,\cdots,n \sim Inv-\chi^2\left(n-1,\frac{n}{n-1}s^2\right)$$

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

(12) $$\theta_i \mid y_i,\mathbf{x}_i,\boldsymbol{\beta},\sigma^2,i=1,\cdots,n \sim N\left(\frac{4n_iy_i+\frac{\mathbf{x}_i\boldsymbol{\beta}}{\sigma^2}}{4n_i+\frac{1}{\sigma^2}},\frac{1}{4n_i+\frac{1}{\sigma^2}}\right)$$

if Y_i is not missing;

$\theta_i \mid y_i,\mathbf{x}_i,\boldsymbol{\beta},\sigma^2,i=1,\cdots,n \sim N\left(\mathbf{x}_i\boldsymbol{\beta},\sigma^2\right)$ if Y_i is missing.

3 An Example --County-level prevalence of overweight from the 2003 Behavioral Risk Factor Surveillance System

We illustrate the model inference by estimating the county level prevalence of overweight from the 2003 Behavioral Risk Factor Surveillance System (BRFSS). We also propose two ways of detecting outliers and demonstrate them in this application. Bayesian posterior predictive distribution is used to check the model fit under normal and *t* models.

3.1. Data and analysis

Overweight is a risk factor for many diseases including cardiovascular diseases, diabetes, and certain types of cancers. Federal, state, and local government agencies need accurate estimates of prevalence of overweight and obesity for small areas such as counties to implement and evaluate obesity prevention programs. Operationally, an adult is said to be "overweight" if his or her body mass index (BMI) is over 25 where the BMI is defined as weight in kilogram divided by the square of height in meter.

One data source for the estimates of overweight prevalence in small areas is the Behavioral Risk Factor Surveillance System (BRFSS), a telephone survey of the health behaviors of US adults, by the Centers for Disease Control and Prevention (CDC). In 2003, all states and DC used a disproportionate

stratified sample design, a type of list-assisted design (Lepkowski, 1988).

Due to confidentiality concerns, if a county had fewer than 50 subjects in the 2003 public-use BRFSS data, the county identifier was suppressed. In addition, the counties in Alaska are not identified and the whole state of Alaska was treated as one single area ("county"). Totally, 1053 "counties" are identified in the public-use data. Hence there are 1,053 "counties" in the analysis, including the state of Alaska, District of Columbia, and 1,051 counties in the other 49 states. The population sizes of these 1053 "counties" from the 2000 Census ranged from 2,397 to 9,519,338. The average population size is 216,860. The total population in the 1053 "counties" consists 81.1% the total population in US, while the rest 2,061 counties consists 18.9% with a mean population per county 25,748.

The total sample size for the 1053 "counties" is 200,810, excluding those without a valid BMI. The direct estimates of the county level overweight prevalence range from 0.308 to 0.819 with a median 0.604, standard errors from 0.011 to 0.143 with median 0.055.

For every county, we also have four county level variables available. They are obtained from the 2000 Census. These 4 covariates are percentages of Hispanic population, percent of people who have a bachelor or higher degree among those 25 years or over, percent of taking public transportation to work for workers 16 years and over, and the percentages of population that is 0-18 years old. We also considered many other variables related to the county's urban/rural status, MSA status, and various characteristics of the population such as employment status, medium income, poverty level, percent of blue collar workers, etc. They are not contributing much when adjusting for the 4 covariates included in the model. The 4 covariates are included on the log-scale to reduce the impact of skewness and standardized to reduce the impact of collinearity.

We first fit the FH model to the data and identified two outlying counties. A *t* model with unknown degree of freedom *v* is then applied to the data. The parameters in the Gamma prior for *v* is chosen so that it is noninformative. Specifically, $\alpha = \gamma = 10^{-4}$. To obtain the Bayesian estimates under both models, programs are written in GAUSS programming language to run 2000 iterations to each of 10 independent sequences of Gibbs sampler. After the first 1000 iterations in each sequence, i.e., burn-in period, the Gibbs sampler converges according to Gelman-Rubin statistic *R* (Gelman and Rubin, 1992). We then take every 5th draw to avoid autocorrelation between consecutive draws.

3.2 Identify outlying counties under normal model

Under the FH (normal) model, the marginal distribution of y_i by integrating θ_i out is $y_i \sim N(\mathbf{x}_i\boldsymbol{\beta}, \sigma^2 + d_i)$. As suggested in Dempster and Ryan (1985), we have $\delta_i(\sigma^2, \boldsymbol{\beta}) \equiv \frac{y_i - \mathbf{x}_i\boldsymbol{\beta}}{\sqrt{\sigma^2 + d_i}} \sim N(0,1)$ where $\delta_i^2(\sigma^2, \boldsymbol{\beta})$ is a Mahalanobis-like distance, as defined in Lange *et al* (1989). When σ^2 and $\boldsymbol{\beta}$ are replaced by the maximum likelihood estimate (MLE) $\tilde{\sigma}^2$ and $\tilde{\boldsymbol{\beta}}$, $\delta_i(\tilde{\sigma}^2, \tilde{\boldsymbol{\beta}})$ has asymptotically the same $N(0,1)$ distribution. In a Bayesian setting with noninformative prior, the posterior mode is approximately equivalent to the MLE. For the FH model, the posterior mode can further be approximated by the posterior mean $\hat{\sigma}^2$ 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 $y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}$, but also the reliability of the direct estimate y_i. Only those areas with "extreme" and relatively reliable direct estimates (i.e., large $\left|y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}\right|$ and small d_i) will be recognized as outliers.

There is another way to identify outliers. You and Rao (2002) computed statistics proposed by Daniels and Gatsonis (1999). The rationale is that 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 $y_i^{rep} \sim N(\theta_i, d_i)$. The resulting 10,000 draws of y_i^{rep} represents the posterior predictive distribution of y_i.

Define a p-value for the normal model as

$$p_i^N = \frac{1}{J}\sum_{j=1}^{J} I\left[y_i^{rep(j)} > y_i\right]$$
$$I\left[y_i^{rep(j)} > y_i\right] = \begin{cases} 1 & \text{if } y_i^{rep(j)} > y_i \\ 0 & \text{otherwise} \end{cases} \quad (12)$$

where $j = 1, \cdots, J$ 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 $p_i^N < 0.05$ or $p_i^N > 0.95$ is an outlier.

Figure 1 shows the half-normal plot of $\delta_i(\hat{\sigma}^2, \hat{\boldsymbol{\beta}})$ under the FH (normal) model. Two counties at the lower left corner show a violation of normal distribution assumption. Both outlying counties have a $\delta_i(\hat{\sigma}^2, \hat{\boldsymbol{\beta}})$ smaller than -3.5. Both counties have a p-value p_i^N greater than 0.95.

Table 1 gives the direct estimate, posterior mean from the normal model, 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, Dickinson County, Kansas and Park County, Montana. The two counties have direct estimates 0.3-0.4 with a small standard error.

Table 1: 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

Note: County A: Dickinson County, KS, County B: Park County, MT

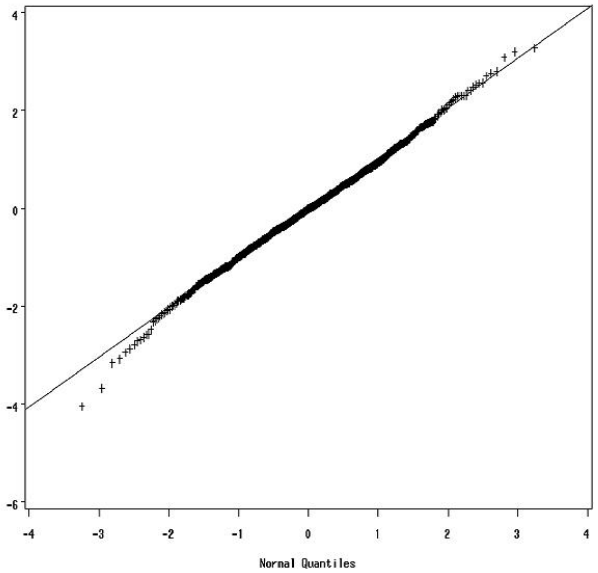
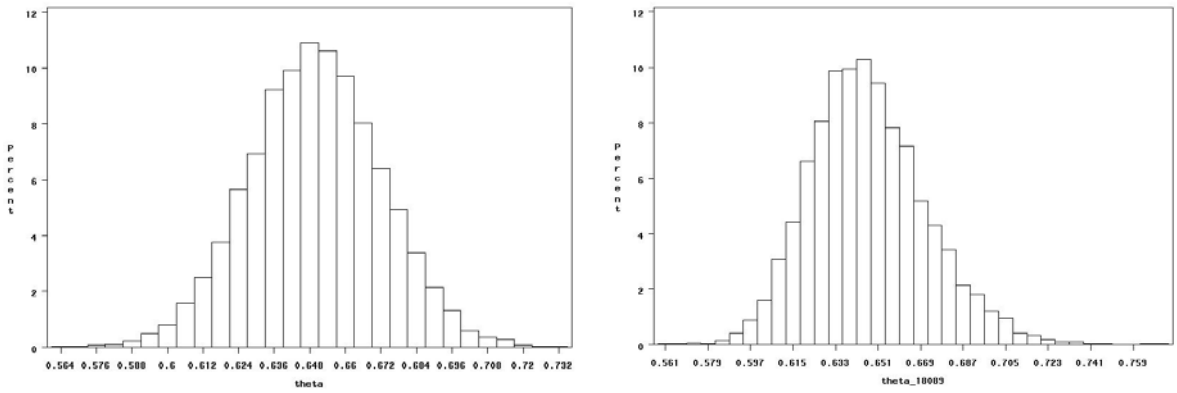


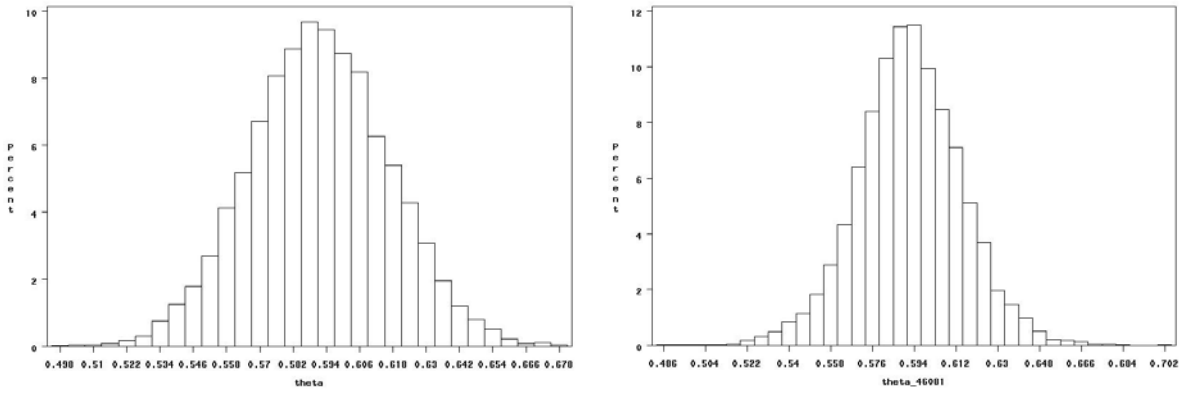
Figure 1: Half-normal plot of $\delta_i(\hat{\sigma}^2, \hat{\boldsymbol{\beta}})$ under FH (normal) model. The straight line is the expected line when there are no outliers.

3.3 Small area estimates

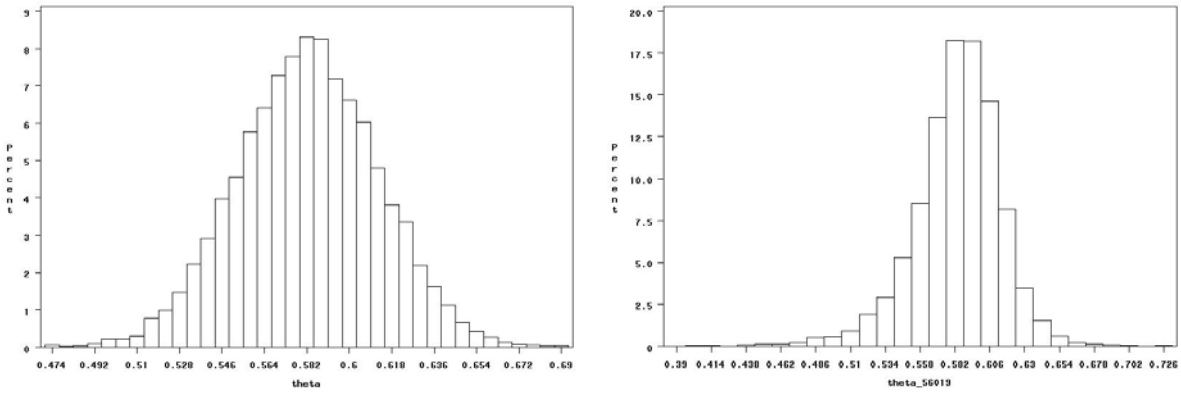
We then compare the posterior distribution of θ_i and one striking difference is in the skewness. In the normal model the posterior distribution of θ_i is approximately symmetric, while in the *t* model it is skewed except for those with a p_i^N close to 0.5. Figure 2 shows histograms from selected counties (including the two outlying counties) with p_i^N from 0.3 to 0.962. We observe that the distribution of θ_i is left skewed when p_i^N <0.5 and right skewed when p_i^N >0.5. Generally the skewness is associated with the distance between p_i^N and 0.5 with the only exception of the distribution of θ_i for the most outlying county with p_i^N =0.962 (Figure 2e). This case needs further exploration.



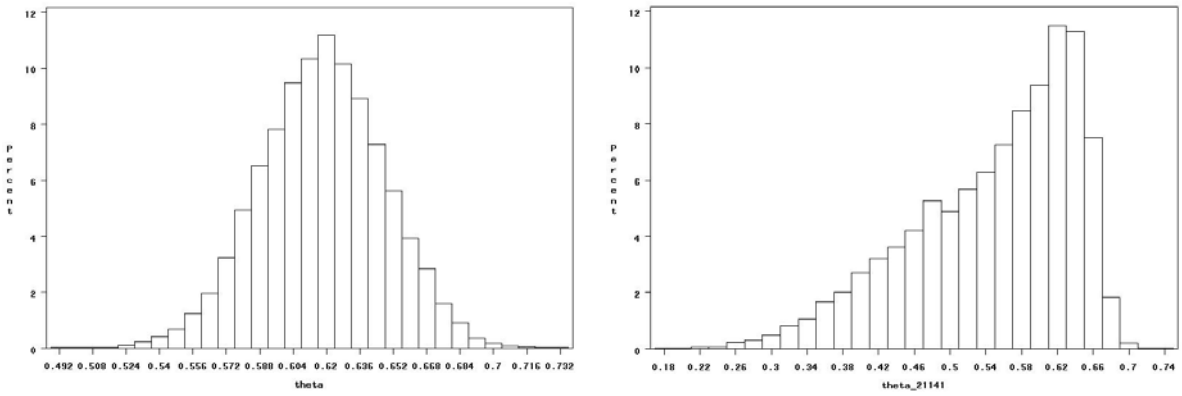
a. posterior distribution of θ_i from the normal model and t model for a county with p_i^N =0.3



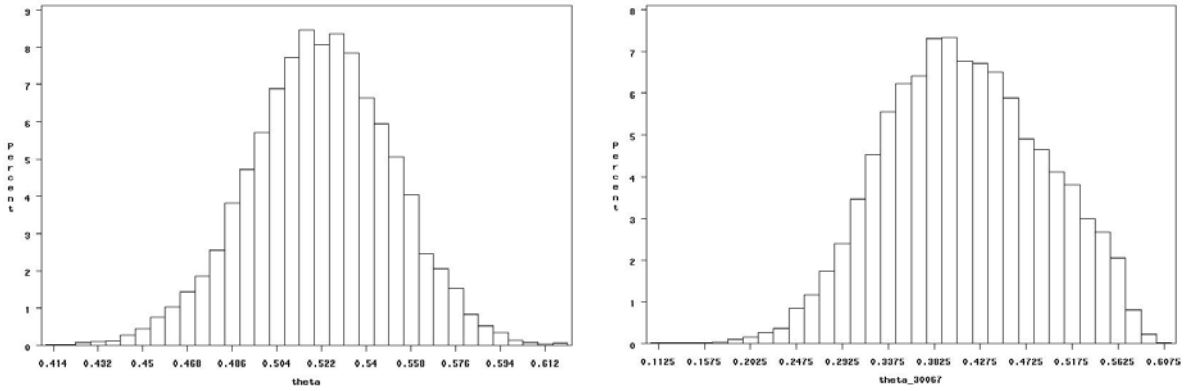
b. posterior distribution of θ_i from the normal model and t model for a county with p_i^N =0.5



c posterior distribution of θ_i from the normal model and t model for a county with p_i^N =0.7



d. posterior distribution of θ_i from the normal model and t model for a county with p_i^N =0.952



e. posterior distribution of θ_i from the normal model and t model for a county with p_i^N =0.962

Figure 2 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)

3.3.1 Posterior means

Figure 3a shows the posterior mean of θ_i from the normal 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 (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 3b.

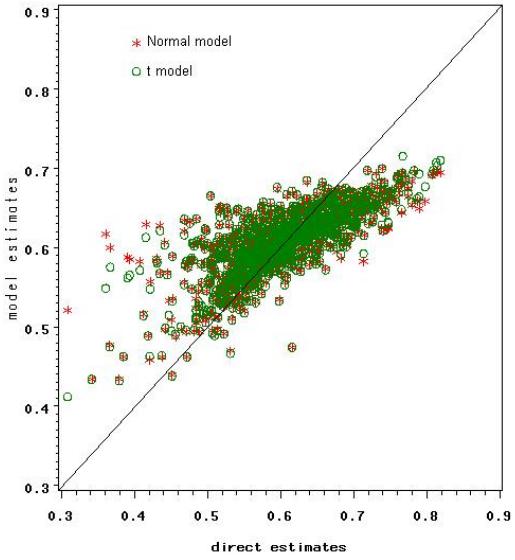


Figure 3a. Posterior mean $\hat{\theta}_i$ from the normal and *t* model versus direct estimates. The line is what to expect if there is no shrinkage effect. The stars are posterior mean from the normal model, the open circles are from the *t* model.

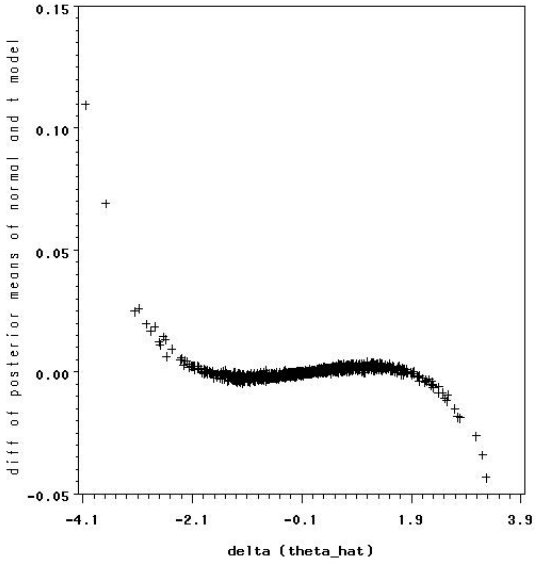


Figure 3b. Differences between the posterior mean $\hat{\theta}_i$ from the normal and *t* model and $\delta_i(\hat{\sigma}^2, \hat{\boldsymbol{\beta}})$. The two points on the left upper corner correspond to the two outlying counties.

3.3.2 Posterior standard derivations

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 2 shows that the posterior standard deviations for the two outlying counties are larger under the *t* model than those for the normal under or even the direct estimates. This is expected due to the high variation in the draws on the tails of a *t* distribution. Although the normal model estimates appear to have a higher precision in these areas, the model does not fit the data well. To check the model fit for the *t* model, we define the same p_i^t for the *t* model as in (12). The values of p_i^t for all counties are between 0.166 and 0.879.

3.4 Estimate of *v*

The posterior distribution of *v* is shown in Figure 4. The posterior mean of *v* is 3.96 with a standard deviation 0.16 with a 95% posterior confidence interval (3.62, 4.24). Although it appears a little right-skewed, the mode of *v* is very close to the posterior mean 4.00.

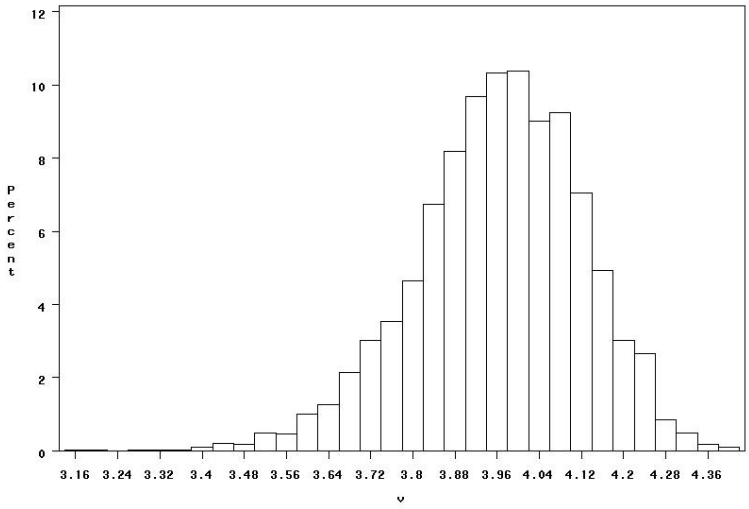


Figure 4: Histogram of draws of degree of freedom *v* from its posterior distribution after burn-in period in Gibbs sampling

3.5 Estimates of σ^2 and $\boldsymbol{\beta}$

Table 2 gives the estimates of σ^2 and $\boldsymbol{\beta}$. The posterior mean and standard deviation of $\boldsymbol{\beta}$ from the two models do not differ much. From both models, we can see that percent of bachelor degree or higher education in 25+ and percent of population 0-18 years old explain more variation in the county level prevalence of overweight than other county covariates. The prevalence of overweight is positively correlated with percent of population 0-18 years, and negatively correlated with percent of bachelor degree or higher education in people 25 years old or over, percent of Hispanic population, and percent of taking public transportation to work in workers 16 years old or over.

Table 2 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
$\boldsymbol{\beta}$	% of Hispanic	-0.0065 (0.0022)	-0.0070 (0.0025)
	% of taking public transportation to work in workers 16+	-0.0060 (0.0022)	-0.0060 (0.0022)
	% of population 0-18 yr old	0.0144 (0.0019)	0.0148 (0.0021)
	% of bachelor degree or higher education in 25+	-0.0278 (0.0018)	-0.0276 (0.0019)
σ^2		0.0010 (0.0001)	0.0005 (0.0001)

The interpretations of σ^2 are different in the normal and *t* models. It is not meaningful to compare σ^2 under different models. However, when we discuss the weight w_i in Section 2.1, we assume that σ^2 and $\boldsymbol{\beta}$ are known constants. Note that this assumption does not hold, especially for σ^2.

4 Discussion

In this article we illustrates the ability of improving model fit using a *t* distribution to model small area mean where there are outliers. We conclude with some remarks on the limitations of the approach, and mention some areas that seem to require further study.

1. Our model assumes an independent sampling error among areas. Due to complex nature of the sampling designs, the sampling errors are sometimes correlated. For example, when the small areas cut across the sampling design. Under this situation, we need to consider a multivariate model for y_i, i.e., $\mathbf{y} \sim N(\boldsymbol{\theta}, \mathbf{D})$

where $\mathbf{y} = (y_1, \cdots, y_k)'$, $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_k)'$, $\mathbf{D}$ is the sampling variance matrix of $\mathbf{y}$. It is the case discussed in Datta and Lahiri (1995). Under Bayesian framework the extension is straightforward. With the model for $\boldsymbol{\theta}$ unchanged, the conditional distributions of all parameters are unchanged except for $\boldsymbol{\theta}$.

2. We assume a normal distribution 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 complex sampling design, one can assume a binary model for $n_i y_i$, i.e., $n_i y_i \sim Binomial(n_i, \theta_i)$, where n_i is the sample size from area i . We can further assume $\text{logit}(\theta_i) \sim t_v(\mathbf{x}_i\boldsymbol{\beta}, \sigma^2)$. Rao (2003) discussed the normal case where $\text{logit}(\theta_i) \sim N(\mathbf{x}_i\boldsymbol{\beta}, \sigma^2)$ 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 $n_i y_i \sim Bin(n_i, \theta_i)$.

3. To detect outliers, we defined $\delta_i(\hat{\sigma}^2, \hat{\boldsymbol{\beta}})$ as a pseudo measure in Section 3.2. The distribution of $\delta_i(\hat{\sigma}^2, \hat{\boldsymbol{\beta}})$ is approximated by assuming the posterior mean of σ^2 and $\boldsymbol{\beta}$ under noninformative prior are roughly equal to the MLE of σ^2 and $\boldsymbol{\beta}$. There might be some bias in the approximation. The posterior predictive distribution approach is more appropriate in general.

4. One can also consider proper priors for σ^2 in the normal and *t* models. A typical choice for the prior of σ^2 for the normal model is inverse Gamma, and for the *t* model, Gamma. When both priors are chosen to be noninformative, little difference is found between the posterior distributions under proper and improper priors.

REFERENCES

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm, *The American Statistician*, **49** , 327-335

Daniels, M.J. and Gatsonis, C. (1999), Hierarchical generalized linear models in the analysis of variation in heath care utilization, Journal *of the American Statistical Association*, **94**, 29-42

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

Dempster, A. P., and Ryan, L. M. (1985). Weighted normal plots, *Journal of the American Statistical Association*, **80**, 845-850

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

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian data analysis*, London and New York: Chapman & Hall

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (Disc: p483-501, 503-511), *Statistical Science*, **7** , 457-472

Gelfand, A. E. , and Smith, A. F. M. (1991), Gibbs sampling for marginal posterior expectations, *Communications in Statistics, Part A -- Theory and Methods*, **20**, 1747-1766

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

Lange, K. L., Little, R. J. A., and Taylor, J. M. G. (1989). Robust statistical modeling using the *t*-distribution, *Journal of the American Statistical Association*, **84** , 881-896

Lepkowski, J. M. (1988), Telephone sampling methods in the United States, *Telephone Survey Methodology*, 73-98

Pinheiro, J. C., Liu, C., and Wu, Y. N. (2001). Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate *t* distribution, *Journal of Computational and Graphical Statistics*, **10 (2)** , 249-276

Prasad, N. G. N. and Rao, J. N. K. (1990). The estimation of the mean squared error of small-area estimators, *Journal of the American Statistical Association*, **85**, 163-171

Raghunathan, T. E. , and Rubin, D. B. (1990). An application of Bayesian statistics using

sampling/importance resampling for a deceptively simple problem in quality control, *Data Quality Control: Theory and Pragmatics*, 229-243

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

Tierney, L. (1991). Exploring posterior distributions using Markov chains, *Computing Science and Statistics. Proceedings of the 23rd Symposium on the Interface*, 563-570

Watanabe, Toshiaki (2001), ``On sampling the degree-of-freedom of Student's-t disturbances, *Statistics & Probability Letters*, **52 (2)** , 177-181

You, Y. and Rao, J. N. K. (2002b), small area estimation using unmatched sampling and linking models, *Canadian Journal of Statistics*, **30**, 3-15