Nonlinear mixed effects cross-sectional models for estimation of smoking proportions using the National Health Interview Survey

Neung Soo Ha* and Van Parsons*

*National Center for Health Statistics, 3311 Toledo Rd, Hyattsville, MD 20782

Abstract

The National Health Interview Survey (NHIS) is an annual health survey conducted by the National Center for Health Statistic. The survey design provides reliable annual estimates for health-related conditions for the nation and the four major geographical regions of the United States. However, direct estimates for some states or sub-state regions are unreliable due to insufficient sample sizes. We propose small area models that include both local area and short-term time random effects, which describe year-to-year variation over cluster samples, to estimate the prevalence of smoking for each of the fifty U.S. states and the District of Columbia. In particular, hierarchical Bayesian nonlinear mixed effect models, using the 2006 to 2010 NHIS data, is explored. Auxiliary variables will be obtained from the Area Resource File. Bayesian Markov Chain Monte Carlo (MCMC) approaches will be used for estimation. A major portion of this study is a discussion of various methods to estimate the time specific sampling covariances needed to implement the proposed models. Comparison of different models by model fits and model performance are discussed.

Key Words: Hierarchical Bayesian modeling, MCMC, small area estimation, cross-sectional time-series model, National Health Interview Survey.

1. Introduction

Most large-scale sample surveys are designed to provide reliable estimates for large geographical regions and large subgroups of a population. The National Health Interview Survey, for example, is an annual health survey with complex design that provides reliable annual estimates on health related topics, such as insurance coverage and smoking rates for the nation and four Census regions: Northeast, South, Midwest, West. In many instances, however, estimates for smaller regions or cross sectional domains are also needed for formulating government policies.

Because direct survey estimates, based solely on sampled units for small area or small domains, are likely to yield unreliable values due to small sample sizes and sampling strategies, analysts use a number of small area estimation (SAE) techniques that 'borrow strength' through using implicit or explicit models. These models utilize a link between small areas and other supplementary data, such as administrative records or values from other surveys, (e.g. the American Community Survey, (`http://www.census.gov/acs/www/`)). Rao (2003) has a thorough review on various models for small area estimations.

Most of the SAE research is focused on cross-sectional data at a given point in time, but there are a few papers that use time series methods, such as Scott and Smith (1974), Jones (1980), and Binder and Dick (1989), but their methods have failed to combine time series and cross-sectional data. The main purpose of this paper is to propose a cross-sectional and

[1] *"The findings and conclusions in this paper are those of the authors and do not necessarily represent the views of the National Center for Health Statistics, Centers for Disease Control and Prevention."*

time series model that uses autocorrelated random effects and a known sampling covariance over time. The true value of the covariance matrix is unknown; thus in practice, its smoothed estimate is used instead. There have been papers that propose different methods for smoothing the matrix. Datta et al. (1999) have explored a model with a long time series, (48 months), and You et al. (2003) have applied their model with reduction of the coefficient of variation (CV). In this paper, we present alternative models with an emphasis on techniques for smoothing the sampling covariance matrix and evaluate their performances against each other.

This paper is organized as follows: In section 2, we review various small area models proposed in the literature. In Section 3, we discuss our methods of smoothing the sampling covariance matrix and discuss parameter estimation methods from a non-linear mixed effects model via using MCMC techniques. In Section 4, we examine our results from using the 2006 to 2010 NHIS data. In section 5, we provide our results and then conclude by offering final comments and future work directions.

2. Review of Small Area Models

In general, small area estimation using area level models can be implemented by combining direct area level estimates and auxiliary information in a two level modeling. Fay and Herriot (1979) demonstrated this property in the following way. Let $\hat{\theta}_{it}$ be the observed response, θ_{it} be the parameter of interest, (e.g. total, mean, or quantile), x_{it} be the covariates observed or recorded for area i, (, $i = 1, \ldots, m$, where m is the total number of small areas), and $t = 1, \ldots, T$ be the survey years. The Fay-Herriot model is defined as:

$$\begin{aligned} &\textit{L1}\text{: Sampling model } :\hat{\theta}_{it}|\theta_{it} \overset{ind}{\sim} \mathcal{N}(\theta_{it}, \sigma^2_{it}), \\ &\textit{L2}\text{: Linking model } :\theta_{it} \overset{ind}{\sim} \mathcal{N}(\boldsymbol{x}'_{it}\boldsymbol{\beta}, \sigma^2_{\nu}), i = 1, \ldots, m,\ t = 1, \ldots, T \end{aligned}$$

where σ^2_{ν} is the variance of the time area-specific random effect and the link is the identity function. The Fay-Herrriot model is used to obtain cross-sectional estimates, but this model fails to include year-to-year variability within a given area because the covariance structure in the sampling model is zero for different time periods, i.e. $Cov(\hat{\theta}_{it}, \hat{\theta}_{is}) = 0,\ t \neq s$.

2.1 Cross-sectional and time series model

In the NHIS design, primary sampling units (PSUs) are usually counties or contiguous counties. Given a sample design, the same PSUs are visited for all years under that design, but within each PSU, samples of different households are taken. For that reason, there is a geographical overlap between samples even though there is no sample overlap from year to year. Thus, the covariance structure between year-to-year sampling errors ($Cov(\hat{\theta}_{it}, \hat{\theta}_{is}), t \neq s$) has to be considered in the model building.

With the covariance structure, the Fay-Herriot model can be adjusted in the following way. Let $\hat{\boldsymbol{\theta}}_i = (\hat{\theta}_{i1}, \ldots, \hat{\theta}_{iT})', \boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{iT})', \boldsymbol{e}_i = (e_{i1}, \ldots, e_{iT})'$. Assume that $\boldsymbol{e}_i \sim \mathcal{N}(\mathbf{0}, \Sigma_i)$, then the Fay-Herriot with cross sectional time-series model becomes:

$$\begin{aligned} &L1 : \hat{\boldsymbol{\theta}}_i \overset{ind}{\sim} \mathcal{N}(\boldsymbol{\theta}_i, \Sigma_i) \\ &L2 : \theta_{it} = \boldsymbol{x}'_{it}\boldsymbol{\beta} + \nu_i + u_{it},\ \ i = 1, \ldots, m; t = 1, \ldots, T \end{aligned} \tag{1}$$

The linking model in equation (1) needs closer inspection. This model contains a time-specific area-level auxiliary variable, $\boldsymbol{x}'_{it}$, and two random effects: a time-random process, u_{it}, and an area-random effect, $\nu_i \sim N(0, \sigma^2_\nu)$. These random effects can be regarded as separate error sources. Generally, the area-random effect has the property, $\nu_i \sim N(0, \sigma^2_\nu)$. For time-random effect, You et al. (2003) have suggested a random walk model:

$$u_{it} = u_{i,t-1} + \epsilon_{it},$$

where $\epsilon_{it} \overset{ind}{\sim} \mathcal{N}(0, \sigma^2_\epsilon)$, and $\text{cov}(u_{it}, u_{is}) = \min(t, s)\sigma^2_\epsilon$.

Combining the two levels in equation (1), we obtain a linear mixed model with the time component as:

$$\hat{\theta}_{it} = \boldsymbol{x}'_{it}\boldsymbol{\beta} + \nu_i + u_{it} + e_{it}, \;\; i = 1, \ldots, m, t = 1, \ldots, T. \tag{2}$$

However, equation (2) is limited when the parameter of interest is a proportion because direct application of this model could result in negative estimates. For this project, we will use the logit link for equation (1) as in Liu et al. (2007):

$$L2 : \text{logit}(\theta_{it}) = \boldsymbol{x}'_{it}\boldsymbol{\beta} + \nu_i + u_{it}, \;\; i = 1, \ldots, m; t = 1, \ldots, T. \tag{3}$$

3. Models: Smoothing Methods For Sampling Variance

In area-level models of SAE, the smoothed estimates are used for known sampling variances in model constructions, Rao (2003). In this paper, we propose several models with different techniques for smoothing the components of the sampling variance covariance matrix.

Model 1
Our first method uses the generalized variance function (GVF). A GVF is a model that describes the relationship between a statistic and its corresponding variance, and it has traditionally been used for calculating variance estimates, Otto and Bell (1995). Our covariance matrix, Σ_i, is then defined as:

$$\hat{\Sigma}_i = \begin{pmatrix} \hat{\sigma}^2_{i1} & 0 & \cdots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & \cdots & \hat{\sigma}^2_{iT}, \end{pmatrix}$$

i.e. $\hat{\Sigma}_i = \; diag\{\hat{\sigma}^2_{it}\}^T_{t=1}$, and $\hat{\sigma}^2_{it}$ is a variance estimate from a fitted GVF model. We will use this method as a base case and make comparisons with other methods.

For the second model, we keep the same covariance structure as in method 1, but the diagonal terms, $\hat{\sigma}^2_{it}$, are estimated by decomposing the true sampling variance into variance estimate of the sampling proportion under simple random sampling (SRS) multiplied by the design effect estimate over pooled data covering the nine Census divisions. Thus, our variance estimate becomes:

Model 2

$$\hat{\Sigma}_i = \ diag\{\hat{\sigma}_{it}^2\}_{t=1}^T$$
$$\hat{\sigma}_{it} = \frac{\hat{\theta}_{jt}(1-\hat{\theta}_{jt})}{n_{it}} deff_{jt}^{div},$$
$$j = 1, \ldots, 9, \ \ i = 1, \ldots, 51, \ t = 1, \ldots 5,$$

where $\hat{\theta}_{jt}$ is the estimate for Census division j which contains state i, n_{it} = state sample size, and $deff_{jt}^{div}$ is the Census divisional design effect estimate at time t.

For the third model, we use the model similar to that of You (2008). This model is defined as:
Model 3

$$\begin{aligned} Level1(\text{sampling model}) \quad &: \quad \hat{\boldsymbol{\theta}}_i | \boldsymbol{\theta}_i, \hat{\Sigma}_i \overset{\text{ind}}{\sim} \mathcal{N}(\theta_i, \hat{\Sigma}_i), \\ Level2(\text{linking model}) \quad &: \quad logit(\theta_{it}) = \mathbf{x}_{it}' \boldsymbol{\beta} + \alpha_i + u_{it}, \\ \text{Random Walk} \quad &: \quad u_{it} = u_{i,t-1} + \epsilon_{it}, \end{aligned}$$

where $\hat{\Sigma}_i$ is defined as:

$$\hat{\Sigma}_i = \begin{pmatrix} \hat{\sigma}_{i1}^2 & & \hat{\sigma}_{its} \\ & \ddots & \\ \hat{\sigma}_{its} & & \hat{\sigma}_{iT}^2, \end{pmatrix}$$

where $\hat{\sigma}_{it}^2 = \hat{\theta}_{it}(1-\hat{\theta}_{it}) \cdot \frac{1}{T} \sum_{t=1}^T \frac{deff_{jt}^{div}}{n_{it}}$, and $\hat{\sigma}_{itt'} = \hat{\rho}_{|t-s|} \hat{\sigma}_{it} \hat{\sigma}_{is}$. The correlation parameter $\hat{\rho}_{|t-s|}$ is the average year-to-year correlation coefficient. Note that $\hat{\rho}_{|t-s|}$ only depends on time difference.

In order to estimate $\hat{\rho}_{|t-s|}$, You (2008) has used correlation monthly coefficient estimates from previous year surveys. In our case, we cannot use the similar method because past designs of NHIS are different from the current survey's design. Instead, we have made the following assumptions. First, the year-to-year design remains relatively unchanged so that the design effects vary only a little over the years. Second, variation of variance estimates due to different sample sizes would remain small between years. Third, the state-wide year-to-year covariance is similar to that of the entire country. Since there are four correlations covering 2006-2010, $\hat{\rho}_{|t-s|}$ is calculated by averaging over different year intervals using the design-based estimates. For example, parameter $\hat{\rho}_1$ is calculated by averaging over all one year difference intervals, and $\hat{\rho}_2$ is calculated by averaging over two year intervals, and so on. Thus, by using the smoothing method and calculating correlation coefficients, the covariance matrix, $\hat{\Sigma}_i$, can be obtained.

3.1 Inference on Small Area Estimation

For inference about parameters, we have used hierarchical Bayesian (HB) approach by using the computer package, WinBUGS. The WinBUGS is a computer software package that fits the the HB models using the MCMC method, then the Bayesian inference can be

derived based on the MCMC techniques. Following a suggestion by Gelman (2006), we chose uniform prior distributions for σ_ν and weakly informative prior distributions, normal distribution with wide variance, for $\boldsymbol{\beta}$'s.

For convergence diagnostics we have calculated the Gelman and Rubin statistic $\hat{R}$,(Gelman et al. (2004)), from three parallel chains, checked auto-correlation function (ACF) and trace plots for all parameters. Each chain has $100,000$ iterations with burn-in of $50,000$, and thinning is done at 10 iterations. For more information about different convergence diagnostics for MCMC methods, see Cowles and Carlin (1996).

3.2 Selection of auxiliary variables

For our area-specific auxiliary variables, we have used data from the Area Resource File (ARF). More information about the ARF can be found at `http://arf.hrsa.gov/faqs.htm`. We obtained state and time specific auxiliary variables by appropriate aggregation. For all three models, we have used identical auxiliary variables: unemployment rate, minority population rate, senior (aged 65+) population rate, and poverty rate. We have chosen these variables based on standard regression model selection techniques and have used logistic regression using data from fifteen largest states for each year. We have found that these covariates significant in model analysis.

4. Comparison of different models

4.1 Calibration Diagnostic

One of the possible deficiencies of model-based small area estimates occurs when an estimate is aggregated for a larger geographical area and the aggregation is quite different than that of the corresponding direct design estimate. Since direct design based estimates at higher levels of aggregations have better properties of design consistency, analysts tend to prefer models that produce estimates which are closer to the direct estimates under appropriate aggregation. For our analysis, we have compared posterior estimates from each model at the aggregate level from each year for model comparison. We examined the following relative error, RE, at the Census regions:

$$\text{RE } = \left| \frac{\sum_{i \in j} \hat{N}_{it} \hat{\theta}^{ps}_{it} - \hat{\theta}^{dsgn}_{jt}}{\hat{\theta}^{dsgn}_{jt}} \right|, j = 1, \ldots, 4,$$

where $\hat{\theta}^{ps}_{it}$ = posterior mean from the HB model from state i at time t, $\hat{N}_{it} = \sum_{k \in i} w_{ikt}$, w_{ikt} = survey weight for individual k in state i, and $\hat{\theta}^{dsgn}_{jt}$ is the direct estimate for Census region j at time t. The following table displays the result from each model for each year. From Table 1, we can see that all models performed well in region 4. However, the RE does not show clearly which model produces the best overall performance. For example in 2008, model 2 is the best overall while in 2007, model 3 is the best.

4.2 Coverage Diagnostic

This diagnostic evaluates the validity of the confidence intervals generated by the model-based predicted values. It assumes that if the model is correct, the variability of generated estimates would be similar to the variability of the observed values 95% of the time, Gelman

2006		models	
cen_rgn	1	2	3
NE	0.0514	0.0340	0.0026
S	0.0598	0.0516	0.0731
MW	0.0243	0.0049	0.0460
W	0.0579	0.0129	0.0323
2008		models	
cen_rgn	1	2	3
NE	0.0140	0.0016	0.0005
S	0.0144	0.0055	0.0313
MW	0.0542	0.0361	0.0527
W	0.1001	0.0622	0.0660
2010		models	
cen_rgn	1	2	3
NE	0.0399	0.0687	0.0233
S	0.0185	0.0447	0.0018
MW	0.0135	0.0165	0.0162
W	0.0006	0.0527	0.0126

2007		models	
cen_rgn	1	2	3
NE	0.0155	0.0343	0.0093
S	0.0186	0.0278	0.0163
MW	0.0112	0.0094	0.0339
W	0.0348	0.0079	0.0068
2009		models	
cen_rgn	1	2	3
NE	0.0653	0.0612	0.0260
S	0.0188	0.0160	0.0262
MW	0.0306	0.0206	0.0264
W	0.0092	0.0204	0.0018

Table 1: Annual RE, NE=Northeast, S=South, MW=Midwest, W=West

et al. (2004). Let $\boldsymbol{y}_{obs}$ denote the observed data and $\boldsymbol{y}_{new}$ be generated data from a posterior predictive distribution, $f((\boldsymbol{y}, \boldsymbol{\theta})|\boldsymbol{y}_{obs})$. Then, we define our sample variabilities as:

$$S_{0,t} = \sum_{i=1}^{51} (y_{i,t} - \bar{y}_t)^2$$

$$S_{(\ell),t} = \sum_{i=1}^{51} (y_{i,t}^{(\ell)} - \bar{y}_t^{(\ell)})^2, \ \ell = 1, \ldots, 15000,$$

where $S_{0,t}$ denotes observed sample variability, $S_{(\ell),t}$ denotes predicted sample variability, ℓ = number of iterations, and t = survey year. We want to check if the 95% posterior predictive interval, defined as $P(a < S_{(\ell),t} < b) = .95$ with a pair of numbers (a, b), contains $S_{0,t}$.

Year	$S_{0,t}$	1	2	3
2006	0.1463	(0.1461, 0.3289)	(0.1028, 0.2176)	(0.0998, 0.2291)
2007	0.1616	(0.1260, 0.2890)	(0.1181, 0.2606)	(0.1112, 0.2483)
2008	0.1368	(0.3700, 1.2627)	(0.1219, 0.2684)	(0.1154, 0.2502)
2009	0.2118	(0.1054, 0.1929)	(0.1098, 0.2391)	(0.1278, 0.2813)
2010	0.1578	(0.1004, 0.1853)	(0.1020, 0.2141)	(0.1318, 0.2909)

Table 2: 95% coverage

From Table 2, we can see that 95% interval created by predicted values from model 1 fails to include the observed value for 2008 and 2009 while models 2 and 3 succeeded; model 1 is inferior according to this diagnostic.

4.3 Model comparison under the posterior predictive divergence approach

To compare different methods, we computed the Laud-Ibrahim divergence measure, Laud et al. (1995), which is given by:

$$d(\boldsymbol{y}_{new}, \boldsymbol{y}_{obs}) = E(n^{-1}\|\boldsymbol{y}_{new} - \boldsymbol{y}_{obs}\|^2|\boldsymbol{y}_{obs}),$$

where n is the dimension of $\boldsymbol{y}_{obs}$. This divergence measure is approximated by $(nB)^{-1}\sum_{\ell}^{B}\|\boldsymbol{y}^{(\ell)} - \boldsymbol{y}_{obs}\|^2$, where n is the dimension of y_{obs} and $B = 15000$. From all models, we prefer the one with the smallest value of this divergence measure, d.

Model	d
1	0.1321
2	0.0894
3	0.0197

Table 3: d = Laud-Ibrahim divergence measure

Table 3 shows that model 3 clearly gives a value that outperforms other models.

5. Estimation

From the previous model comparison analysis, models 1 and 3 appear to emerge as the least and the best models, respectively. We compare model 1 and model 3 methods further.

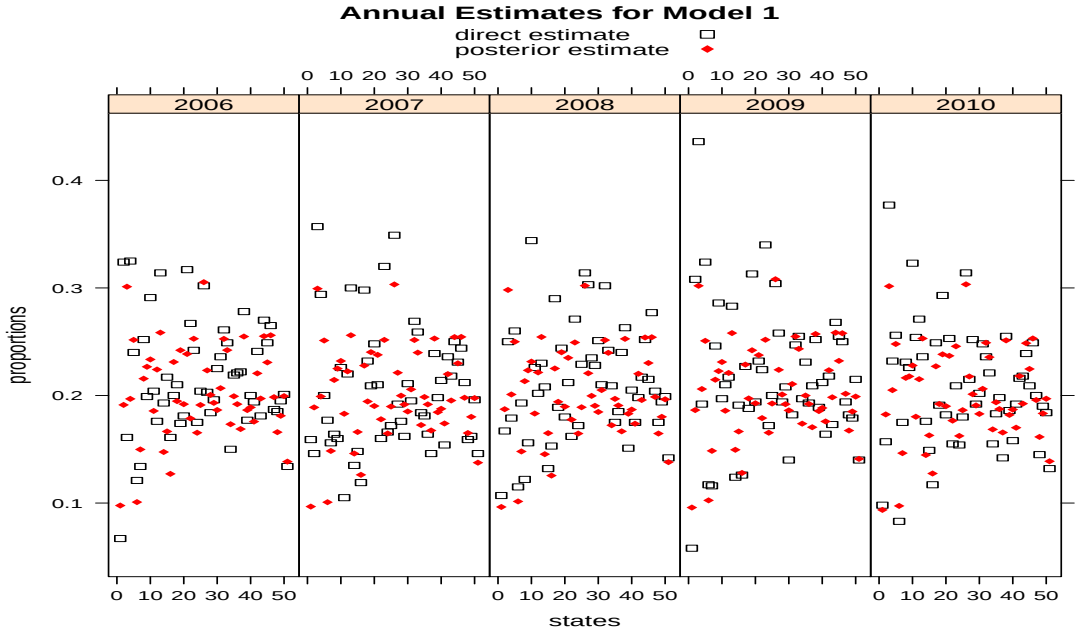

Figure 1: States are arranged from smallest to largest in sample sizes, Posterior estimates show very little variation across different years

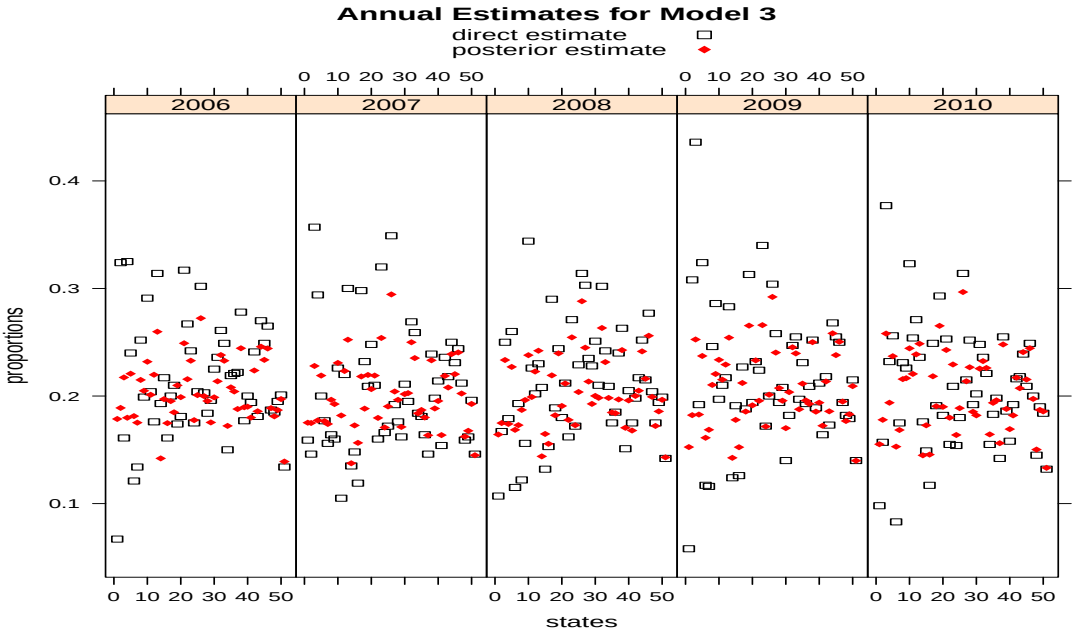

Figure 2: States are arranged from smallest to largest in sample sizes, Posterior estimates show some variation across different years

Figures 1 and 2 show results from models 1 and 3 respectively. For confidentiality reasons, actual state names are omitted, and states are arranged from smallest to largest in sample sizes. There are similarities from both figures. As we move from the left to the right side of the graph within each year, direct estimates show more variability for smaller states from the modeled estimate, and for larger states, direct estimates are very similar to those of model estimates. However, the difference between model 1 and model 3 is shown by the variability of posterior estimates between years. Model 1 includes sampling variance that has no covariance between different years. As a result, the posterior values between different years show very little changes. On the other hand, posterior estimates from model 3 show more variation between different years. This is a consequence of using the sampling variance matrix in the model. We expect some variation from year-to-year for a given state; thus, model 3 produces more plausible results than model 1.

6. Summary

This study provides estimates for the proportion of smokers from each state by using the NHIS data from 2006 to 2010. The Fay-Herriot type models require knowledge of the true variance at the sampling level but, in practice, variance estimates must be used. Additionally, if there is an assumption about the year-to-year correlation, care must be taken in defining its structure. We have shown three different models with different sampling variance structures and provided their assessment using different diagnostic techniques. Our results have shown that a model that has year to year correlation structure provides more reasonable estimates.

In future research, we would like to explore other models for sub-state estimates, such as state$\times$gender, using the NHIS data and perform additional analysis for model fits and model selections such as Bayesian cross-validation. We hope that this study has provided

some foundation for multi-year data analysis.

References

Binder, D. A. and Dick, J. P. (1989). Modeling and estimation for repeated surveys. *Survey Methodology*, 15:29–45.

Cowles, M. K. and Carlin, B. P. (1996). Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review. *Journal of the American Statistical Association*, 91:883–904.

Datta, G., Lahiri, P., Maiti, T., and Lu, K. L. (1999). Hierarchical Bayes Estimation of Unemplyment Rates for the States of the U.S. *Journal of the American Statistical Association*, 94:1074–1082.

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. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1:515–533.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2004). *Bayesian Data Analysis*. Chapman and Hall/CRC, New York, NY.

Jones, R. (1980). Best linear unbiased estimators for repeated surveys. *Journal of Royal Statistical Society*, 42:221–226.

Laud, P. W., Wisconsin, M. C., and Ibrahim, J. G. (1995). Predictive model selection. *Journal of the Royal Statistical Society, Ser. B*, 57:247–262.

Liu, B., Lahiri, P., and Kalton, G. (2007). Hierarchical Bayes Modeling of Survey-Weighted Small Area Proportions. In *Proceedings of the Survey Research Methods Section, ASA*.

Otto, M. and Bell, W. (1995). Sampling Error Modeling of Poverty and Income Statistics for States. In *Proceedings of the Government Statics Section, ASA*.

Rao, J. N. K. (2003). *Small Area Estimation*. Wiley Series in Survey Methodology, Hoboken, NJ.

Scott, A. and Smith, T. M. P. (1974). Analysis of repeated surveys using time series methods. *Journal of the American Statistical Association*, 69:674–678.

You, Y. (2008). An integrated modeling approach to unemployment rate estimation for sub-provincial areas of Canada. *Survey Methodology*, 34:19–27.

You, Y., Rao, J., and Gambino, J. (2003). Model-based unemployment rate estimation for the Canadian Labour Force Survey: A hierarchical bayes approach. *Survey Methodology*, 29:25–32.