Multi-Level Small Area Estimation using Calibrated Hierarchical Likelihood Approach and Post-Stratification with Bias Correction

by

Nirosha Rathnayake

A DISSERTATION

Presented to the Faculty of
the University of Nebraska Graduate College
in Partial Fulfilment of the Requirements
for the Degree of Doctor of Philosophy

Biostatistics

Under the Supervision of Professor Hongying (Daisy) Dai

University of Nebraska Medical Center
Omaha, Nebraska

May 22, 2020

Supervisory Committee:
Jane Meza, Ph.D.
Kendra Schmid, Ph.D. Steven From, Ph.D.

Small Area Estimation using Calibrated and Post-Stratified Hierarchical (h) Likelihood Approach with Bias Correction

Abstract

Nirosha Rathnayake

University of Nebraska Medical Center,

Supervisor: Hongying (Daisy) Dai, Ph.D.

Small area estimation (SAE) has been widely used in a variety of applications to draw estimates in geographic domains represented as a metropolitan area, district, county, and state. The direct estimation methods provide accurate estimates when the sample size of study participants within each area unit is sufficiently large, but it might not always be realistic to have large sample sizes of study participants when considering small geographical regions. Meanwhile, high dimensional socio-ecological data exist at the community level, providing an opportunity for model-based estimation by incorporating rich auxiliary information at the individual and area levels. Thus, it is critical to develop advanced statistical modeling to extract accurate information when the dimensions of geographic domains and auxiliary variables are increasing, and the sample sizes of study participants within the geographic units are decreasing, even to zero.

The most common SAE modeling technique is the Fay-Herriot (FH) methods based on the unit level and area-level observations, which includes the fixed effects and random effects to account for area-specific variation. Many studies have been conducted in SAE for normally distributed random effects, but this might not be realistic in many scenarios. Here, we focus on a generalization of Henderson's joint likelihood, called a hierarchical or h-likelihood, for inferences in SAE.

The dissertation will cover three aims: Aim 1. Develop a novel modeling approach for SAE via hierarchical generalized linear models. We propose a Calibrated and Post-stratified hierarchical (CPH) likelihood approach to obtain SAE through hierarchical estimation of fixed effects and random effects with bias correction based on Regression Calibration Method (RCM) which is considered in aim 3. Multi-level estimations for different regions are obtained through post-stratification strategy. Unified analysis through the h-likelihood provides flexibility in statistical inferences for unobserved random variables and leads to a single algorithm, expressed as a set of interlinked and augmented GLMs, to be used for fitting broad class of new models with random effects. Aim 2. We will then extend this methodology to joint modeling of multiple outcome

variables with an application of poly tobacco use among adolescents. In aim 3, we will consider measurement error correction in both aim 1 and aim 2. Extensive simulation studies will be conducted to assess the empirical performance of estimation accuracy at varying scenarios. We will illustrate our method using the National Youth Tobacco Survey to assess the effects of tobacco control policy and geographic disparities at the county level. Last, we will develop an R package for SAE modeling using hierarchical likelihood. The asymptotic properties of MHLES are studied.

List of Tables

Table No.	Description	Page No.
3.1	Summary Statistics of prevalence in 143 counties by the total, age group, sex, and race of NYTS 2014-2015 respondents	
3.2	Model estimates using FH model for ever use and current use of E-Cigarettes	

List of Figures

Figure No.	Description	Page No.
3.1	County prevalence of ever use of E-Cigarettes in the U.S. based on 2014-2015 NYTS data	
3.2	County prevalence of current use of E-Cigarettes in the U.S. based on 2014-2015 NYTS data	

List of Abbreviations

Term	Description
$\boldsymbol{X}$	Design matrix for explanatory variables with fixed effects
$\boldsymbol{\beta}$	Regression coefficients for fixed effects
$\boldsymbol{Z}$	Design matrix of explanatory variables with random effects
$\boldsymbol{u}$	Vector of area-specific random effects
$\boldsymbol{m}$	Number of random effects (counties or small areas)
$\boldsymbol{v}$	Random component
$\boldsymbol{y}$	Response vector
$\boldsymbol{n_i}$	Sample size of small area i (i.e., county i)
$\boldsymbol{N}$	Total number of observations
$\boldsymbol{p}$	Number of parameters
$\boldsymbol{T}, \boldsymbol{t}$	Transpose of a matrix or a vector
$\boldsymbol{det}$	Determinant of a matrix
$\boldsymbol{h}$	Hierarchical log likelihood
$\boldsymbol{\ell}$	Log-likelihood of joint density
$\boldsymbol{\ell_R}$	Restricted log-likelihood of joint density
$\boldsymbol{\epsilon}$	Residual or the sampling error
$\boldsymbol{\sigma_e^2}$	Standard deviation of normally distributed sampling error
$\boldsymbol{\sigma_u^2}$	Standard deviation of random effect
$\boldsymbol{G}$	Variance-covariance matrix of random effect $\boldsymbol{u}$
$\boldsymbol{R}$	Variance-covariance matrix of residual $\boldsymbol{\epsilon}$
$\boldsymbol{V}$	Variance-covariance matrix of y

$\boldsymbol{\mu}$	Conditional mean of $\boldsymbol{y}$ given $\boldsymbol{u}$ in GLM
$\boldsymbol{\theta}$	Canonical (natural) parameter in GLM
$\boldsymbol{\phi}$	Dispersion parameter in GLM
$\boldsymbol{g}()$	Link function in GLM
$\boldsymbol{\mathcal{J}}, \boldsymbol{I}$	Fisher or Observed Information matrix
$\boldsymbol{S}$	Gradient vector (or score function)
$\widetilde{\boldsymbol{\beta}}$	Weighted least squares estimator of $\boldsymbol{\beta}$
$\widehat{\boldsymbol{\delta}}$	(Hierarchal) Maximum likelihood estimator of $\boldsymbol{\delta}$
$\widehat{\boldsymbol{\beta}}$	(Hierarchal) Maximum likelihood estimator of $\boldsymbol{\beta}$
$\widetilde{\boldsymbol{\theta}_{\boldsymbol{i}}}$	Best Linear Unbiased Prediction (BLUP) of $\boldsymbol{\theta}_i$

Table of Contents

Chapter 1. Introduction to Small Area Estimation 10
1.1 Background 10
1.2 Direct Estimation Method 12
1.3 Indirect Estimation Method 12
1.3.1 Area-Level Model 13
1.3.2 Basic Unit Level Model 15
1.4 Best Linear Unbiased Prediction (BLUP) and Empirical BLUP (EBLUP) 15
1.5 Model comparison 16
Chapter 2. Advanced Modeling in SAE 18
2.1 Generalized Linear Mixed Model (GLMM) 18
2.1 BLUP and EBULP Estimators 20
2.2 ML and REML Estimators 21
2.3 Empirical and Hierarchical Bayes Method for Small Area Models 23
2.4 Hierarchical Generalized Linear Models for SAE Models 25
2.4.1 Hierarchical Generalized Linear Models 25
2.4.2 H-likelihood and MHLE of $(\boldsymbol{\beta}, \boldsymbol{u})$ 27
2.4.3 Penalized Partial Maximum Likelihood Estimation of dispersion parameters 28
2.5 Literature Review on Advanced Modelling Techniques in Small Area Estimation 31
Chapter 3. Multilevel Small Area Estimation in Survey Research 37
3.1 Emerging tobacco use among adolescents 37
3.2 Materials and Methods 39
3.2.1 GLMM Model Specification 39
3.2.2 County Level Prediction 40
3.4 Results 41
Chapter 4. Small Area Estimation using Calibrated Hierarchical Likelihood Approach and Post-Stratitication 44
4.1 A mixed logistic model for small area estimation 44
4.2 Hierarchical $(\boldsymbol{h})$ - Likelihood 45
4.3 MHLE of $(\boldsymbol{\beta}, \boldsymbol{u})$ 46
4.4 Bias Correction 48
4.5 MHLE of variance parameter $\boldsymbol{\theta}$ 50
4.5.1 Iterative approach based on the partial derivative 50
4.6 Asymptotic Properties of MHLE of $\boldsymbol{\beta}$ 53

4.7 Asymptotic Properties of MHLE of u 53
Chapter 5. Penalized CPH with Bias Correction in small area estimation 56
Chapter 6. Joint modeling of multiple outcomes in small area estimation 61
6.1 Parameter Estimation of Fixed Effects and Random Effects 63
Chapter 7. Simulation 69
7.1 Simulation – Binary HGLM 69
7.1.1 Data Generation 69
7.1.2 MC Simulation Results 70
7.2 Simulation – Joint Modeling 74
Chapter 8. Real Data Analysis 75
8.1 CPH with Bias Correction Approach on Tobacco Smoking Data 75
8.2 Joint Modeling of Multiple Outcomes based on Tobacco Smoking Data 82
Chapter 9. Discussion 82
Appendix 84
Appendix 1. Results for ever use of E-cigarettes based on NYTS data 84
Appendix 2. Results for current use of E-cigarettes based on NYTS data 86
Appendix 3. Simulation Results for Mixed Logit Model Based On CPH with Bias Correction Approach 88
Appendix 4. R program for CPH with Bias Correction Approach in SAE 88

Chapter 1. Introduction to Small Area Estimation

1.1 Background

Small Area Estimation (SAE) is being widely used in recent decades to draw conclusions (estimates) on small areas (for small sample size) mostly in survey sampling studies. Small areas are domains represented as metropolitan area, district, county, or state level. The method has been applied in many cases such as public health related studies, financial assessment, education planning, forest inventory studies, agricultural studies, government (employment and payroll) studies etc. (Zahava Berkowitz et al., 2016; Z. Berkowitz et al., 2019; H. Dai, D. Catley, K. P. Richter, K. Goggin, & E. F. J. P. Ellerbeck, 2018; Fay III & Herriot, 1979; Martuzzi & Elliott, 1996; Mauro, Monleon, Temesgen, & Ford, 2017; Nancy A Rigotti & Sara Kalkhoran, 2017; Yasui, Liu, Benach, & Winget, 2000).

The major challenges of small area estimations in survey sampling occur due to smallness of the sample or no sampled units. The main idea of SAE techniques is to make conclusions on different small geographical areas based on subpopulation data, but when other factors are taken into account (gender, age group, ethnicity, income, education level etc.), and the sample size of certain subpopulations might not be sufficiently large. This occurs mostly because many survey studies only focus on major areas or areas with large population, hence very little or no information is available on small areas. When the sample size is not large enough to extract accurate information from the model, SAE techniques try to provide sufficiently reliable outcomes using advanced statistical modeling. Some of the various statistical modelling techniques in SAE include Symptomatic Accounting Technique (SAT), direct method, synthetic method, and composite method etc. (Arnab, 2017; Gonzalez & Hoza, 1978; Jiang & Lahiri, 2006; John NK Rao & Molina, 2015).

Direct estimation technique provides considerably reliable estimates when the sample size is sufficiently large. However, when small sample sizes occur in small domains, direct estimation gives large standard errors (large coefficient of variation). Indirect estimation methods overcome

this problem by increasing the effective sample size, thus decrease the standard error. Some of the indirect estimation methods include synthetic estimators, composite estimators, and James-Stein estimators, which are not only based on domain information but also on auxiliary population information. James and Stein estimator adjusts the large coefficient of variation due to the smallness of the sample size up to certain extent (Fay III & Herriot, 1979). Although these indirect estimation methods reduce the coefficient of variation, these methods do not consider between-area variation, which could lead to low precision global estimates. Therefore, many researchers consider linear and generalized linear (nonlinear) mixed models by introducing random effects to account for between-area variation and obtaining area-specific random effects and fixed effects simultaneously.

Furthermore, adequate number of research has been done considering both the area and time effects which is not considered in our modelling (Saei & Chambers, 2003). Various types of mixed models have been developed to improve the inference accuracy in small areas (John NK Rao & Molina, 2015). The area (aggregated) level and unit (element) level models are the two major models under this scenario. The area level models are used when the unit level data for auxiliary variables is not available, instead aggregated data is available. Also, some studies are done considering both unit and area level auxiliary variables.

The basic area level model is the Fay Harriot (FH) model, discussed in section 1.3.1, is based on direst survey estimates obtained from unit level data and area level auxiliary data. The basic FH model assumes that the sampling variance (σ_e^2) is known. Many researchers have proposed extensions of basic FH model to improve the direct survey estimates focusing on unknown sampling variance σ_e^2, smoothing of σ_e^2 through generalized variance function (GVF) technique, temporal and spatial correlation effects introduced in spatial FH model (Dick, 1995; Y. You & Chapman, 2006; Y. J. P. o. S. M. S. You, Statistical Society of Canada, 2008). Additionally, extension of basic area level FH model to sub area level modelling is considered to obtain sub area estimators by borrowing strength from sub area level data (Torabi & Rao, 2014).

The following sections in chapter 1 cover the relevant materials and estimation methods. Section 1.2 briefly states the direct estimation technique in small area estimations which is also known as Horvitz-Thompson estimator. Section 1.3 covers the indirect estimation methods, namely, the most common area level FH model and basic unit level model. Model comparison using AIC and BIC is presented in section 1.5 followed by the BLUP and EBLUP of basic area level model which covers in section 1.4. In this paper, we will be using small areas and clusters interchangeably.

1.2 Direct Estimation Method

Direct estimation technique or design based small area estimation provides accurate estimates for small areas with sufficiently large sample sizes, but this method often cannot provide accurate estimates for areas with small sample sizes or no sampled units. The direct estimator of area means $\widehat{\bar{Y}}_i$ when sampling without replacement is obtained using Horvitz-Thompson (H-T) estimator given as

$$\widehat{\bar{Y}}_i = \frac{1}{n_i}\sum_{j\in s_i}\frac{Y_{ij}}{\pi_{ij}} = \frac{1}{n_i}\sum_{j\in s_i} w_{ij}Y_{ij,}$$

where $i = 1, \dots, m$ small areas, Y_{ij} is j^{th} measurement from the sample set s_i in area i, n_i is the sample size for area i, $\pi_{ij} = 1/w_{ij}$ is the probability of selecting j^{th} unit from area i, and w_{ij} is sampling weight.

The direct estimation method is not appropriate in many problems, it provides large variances due to small sample size areas, and also it cannot be used to make conditional inferences unlike in frequentist and Bayesian approaches. Hence, indirect estimation or model based methods are more popular in most research problems.

1.3 Indirect Estimation Method

The direct estimates will provide large standard errors with small areas as it does not increase the effective sample size in small areas, but indirect estimation method takes care of this

issue by increasing the effective sample size, hence will reduce the standard error. Some of the indirect estimators are synthetic estimators, composite estimators, and James-Stein estimators which do not consider between area variation. The most common indirect estimation technique in SAE, Fay-Herriot (FH) model developed by Robert E. Fay III and Roger A. Herriot in 1979, improves the quality of direct estimates by borrowing strengths from the related areas, considering the between area variation, hence it increases the accuracy of the estimates (John NK Rao & Molina, 2015).

1.3.1 Area-Level Model

Area level models are being used when the unit level auxiliary data is not available. The most common area level model is the Fay-Herriot (FH) model, also called the linking model, which is developed by Fay and Herriot by combining two parts: direct area level estimates and the synthetic estimates obtained from linearly related auxiliary model. This is a special case of the linear mixed model. Suppose that, the population is divided into m areas (domains) with sample sizes $n_1, \dots, n_i$ where $i = 1, \dots, m$, then the direct estimate for the variable of interest y_i for area i, $\hat{y}_i$ is given as,

$$\hat{y}_i = \theta_i + e_i \ , \quad e_i \sim N(0, \sigma_e^2) \quad i = 1, \dots, m,$$

where $\theta_i = \bar{y}_i = \sum_{j=1}^{n_i} y_{ij}/n_i \ , i = 1, \dots, m$ and the total sample size, $N = \sum_{i=1}^{m} n_i$.

The synthetic estimates or θ_i is related to area specific auxiliary data $\boldsymbol{X}_i = (X_1, \dots, X_p)$ through a linear model

$$\theta_i = \boldsymbol{X}_i^T \boldsymbol{\beta} + Z_i u_i, i = 1, \dots, m \qquad (1.1)$$

where Z_i's are known positive constants.

Fay Herriot (FH) Model

The FH model estimates balance the bias and precision of the results borrowing the strength of available information. The linking model assumes that the area level true mean θ_i for area i is

linearly related with area level auxiliary variables ($\boldsymbol{X}_i$). To express another way, the area level mean can be represented as a linear combination of fixed and random effects,

$$\theta_i = \boldsymbol{X}_i^T \boldsymbol{\beta} + u_i \tag{1.2}$$

These two parts builds the basic area level model (FH model) as follows.

$$\widehat{y_i} = \boldsymbol{X}_i^T \boldsymbol{\beta} + u_i + e_i, \tag{1.3}$$

$$i = 1, \dots, m, \quad u_i \sim N(0, \sigma_u^2), \quad e_i \sim N(0, \sigma_e^2),$$

where $\boldsymbol{X}_i$ is a vector of covariates for area i, $\boldsymbol{\beta}$ is vector of unknown regression coefficients, u_i is area specific random effect for area $i, u_i \sim N(0, \sigma_u^2)$, assumed to be independent and identically distributed, with unknown σ_u^2, and e_i is sampling error for area $i, e_i \sim N(0, \sigma_e^2)$ with known σ_e^2. $\boldsymbol{X}_i^T \boldsymbol{\beta}$ is the fixed effects. The estimates of $\boldsymbol{\beta}$ and σ_u^2 , denoted by $\widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma_u^2}$, can be obtained using method of moments (MOM), or maximum likelihood (ML) or restricted ML (REML) techniques. However, for known σ_u^2 and unknown $\boldsymbol{\beta}$, the best unbiased predictor for small area mean j, ($\eta_i = \boldsymbol{X}_i^T \boldsymbol{\beta} + u_i$) is given by Best Linear Unbiased Prediction (BLUP) (Henderson, 1950).

The best linear unbiased prediction estimators depend on the variance-covariance of random effects σ_u^2 which can be estimated using ML or REML methods. If both $\boldsymbol{\beta}$ and σ_u^2 in FH model are unknown, the estimate for small area mean is obtained replacing $\boldsymbol{\beta}$ by their estimators $\widehat{\boldsymbol{\beta}}$ and σ_u^2 by $\widehat{\sigma_u^2}$, which are called the Empirical Best Linear Unbiased Prediction (EBLUP) estimators. Similarly, empirical and hierarchical Bayes estimation methods can also provide accurate estimates for small area means (Ghosh & Rao, 1994; Molina & Marhuenda, 2015; John NK Rao & Molina, 2015; Saei & Chambers, 2003; Torabi & Rao, 2014).

Many studies have obtained unit level and area level EBLUPs in SAE applications and have shown that unit level estimates are more accurate than the area level estimates. Similarly, the root mean squared errors (RMSEs) of area level estimates are larger than that of unit level estimates. Furthermore, some studies have shown that RMSEs of direct estimates are comparably larger than RMSEs of EBLUPs estimates (Mauro et al., 2017). The extensions of standard FH model, such as

spatial FH model deals when the auxiliary information is correlated, which is quite often in real applications. Some applications are included in section 2.6 under literature review.

1.3.2 Basic Unit Level Model

The nested error linear regression was first introduced by Fuller and Battesse (1973) is also known as the basic unit level model (Battese & Fuller, 1981; Fuller & Battese, 1973). It assumes that the auxiliary data are available for each element j in each small area i, and also the variable of interest y_{ij} is related through the nested error linear regression model as follows

$$y_{ij} = \boldsymbol{X}_{ij}^T\boldsymbol{\beta} + u_i + e_{ij},$$

$$i = 1, \dots, m, \qquad j = 1, \dots, n_i, \qquad u_i \sim N(0, \sigma_u^2), \qquad e_{ij} \sim N(0, \sigma_e^2),$$

where $\boldsymbol{X}_{ij}^T = \left(X_{ij1}, \dots, X_{ijp}\right)^T$ is auxiliary variables $k = 1, \dots, p$ for each small area i, $u_i \sim N(0, \sigma_u^2)$ is area level random effects, and $e_{ij} \sim N(0, \sigma_e^2)$ is residual and independent of u_i.

1.4 Best Linear Unbiased Prediction (BLUP) and Empirical BLUP (EBLUP)

First, consider the basic area level model described above, from equation (1.3),

$$\hat{y}_i = \boldsymbol{X}_i^T\boldsymbol{\beta} + u_i + e_i, \quad i = 1, \dots, m, \qquad u_i \sim N(0, \sigma_u^2), \qquad e_i \sim N(0, \sigma_e^2).$$

Given that σ_u^2 and σ_e^2 are known, the Best Linear Unbiased Prediction (BLUP) (Henderson, 1950) of θ_i is

$$\tilde{\theta}_i = E(\eta|y) = \boldsymbol{X}_i^T\tilde{\boldsymbol{\beta}} + E(u_i|y) = \boldsymbol{X}_i^T\tilde{\boldsymbol{\beta}} + \tilde{u}_i,$$

where $\tilde{u}_i = \gamma_i(\hat{y}_i - \boldsymbol{X}_i^T\tilde{\boldsymbol{\beta}})$ is the predicted random effect for area i and $\gamma_i = \sigma_u^2/(\sigma_u^2 + \sigma_e^2)$, $\gamma_i \in (0,1)$. The weighted least squares estimator of $\boldsymbol{\beta}$ is

$$\tilde{\boldsymbol{\beta}} = \left[\sum_{i=1}^{m} \frac{1}{\sigma_u^2 + \sigma_e^2}\, \boldsymbol{X}_i\boldsymbol{X}_i^T\right]^{-1} \sum_{i=1}^{m} \frac{1}{\sigma_u^2 + \sigma_e^2}\, \boldsymbol{X}_i\hat{y}_i.$$

When σ_u^2 is unknown, the EBLUP can be calculated by replacing σ_u^2 with a consistent estimator $\hat{\sigma}_u^2$. The EBLUP can be expressed as follows,

$$\tilde{\theta}_i^{\,EBLUP} = \boldsymbol{X}_i^T\tilde{\boldsymbol{\beta}} + \tilde{u}_i.$$

Substituting $\tilde{u}_i = \hat{\gamma}_i(\hat{y}_i - \boldsymbol{X}_i^T\tilde{\boldsymbol{\beta}})$**,** we have

$$\tilde{\theta}_i^{\,EBLUP} = \boldsymbol{X}_i^T\tilde{\boldsymbol{\beta}} + \hat{\gamma}_i\left(\hat{y}_i - \boldsymbol{X}_i^T\tilde{\boldsymbol{\beta}}\right) = \hat{\gamma}_i\hat{y}_i + (1-\hat{\gamma}_i)\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}.$$

Both BLUP and EBLUP can be considered as a combination of direct etimators, $\hat{y}_i$, and regression-synthetic estimators, $\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}$**.** When the measurement error, σ_e^2, is small as compared with the small area random effects, σ_u^2, the direct estimator becomes reliable. Thus, the BLUP and EBLUP estimates stay closer to the direct estimator. In contrast, when the direct estimator is unreliable, the BLUP and EBLUP estimates get closer to the regression-synthetic estimator.

1.5 Model comparison

The model parameters can be obtained using maximum likelihood (ML) and restricted maximum likelihood (REML) techniques, hence the model comparison and the selection are normally done based on common goodness-of-fit measures, including the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). AIC selects the model that produces the closest distribution to the true distribution through asymptotic approximation of Kullback-Liebner information distance (KL divergence). BIC is very similar to AIC except BIC is obtained from Bayesian model comparison through large sample asymptotic approximation of the marginal likelihood (Akaike, 1973; Schwarz, 1978).

$$AIC = -2\ell(A,\boldsymbol{\beta}) + 2(p+1),$$

$$BIC = -2\,\ell(A,\boldsymbol{\beta}) + (p+1)\log N,$$

where p is the number of estimated parameters, $l(\boldsymbol{A},\boldsymbol{\beta})$ is log likelihood, N is the number of observations.

Both AIC and BIC proximity measures are a combination of goodness of fit and model complexity part in terms of number of parameters and number of observations. It provides how

much information is lost when we approximate one distribution with another distribution, hence lower AIC or BIC indicates a better model fit.

Chapter 2. Advanced Modeling in SAE

2.1 Generalized Linear Mixed Model (GLMM)

Linear Mixed Model (LMM) is a powerful and flexible methodology, an extension of linear models by simultaneously analyzing variables with fixed and random effects. The LMM models variety of data types, clustered data, repeated measures, multilevel/hierarchical data, and spatial data with a continuous variable of interest. In many situations, the observations can be continuous, discrete or categorical. Thus, Generalized Linear Mixed Model (GLMM) is an extension of linear mixed modelling by accommodate a broad class of distributions, including both continuous and categorical observations when random effects are considered. The random effects are considered when observations are correlated within the group or between groups, hence GLMM do not require the assumption of independency of observations as in linear models, but the mean of outcome is linearly related with auxiliary data through a link function depending on observation type (binary, continuous, or count), and the variance is a function of the mean. GLMM models a variety of exponential family distributions such as normal, binomial, Poisson, exponential, multinomial etc (McCulloch & Searle, 2004; Schwarz, 1978). The basic univariate unit-level models in small area estimation are special cases of general linear mixed models (Datta & Ghosh, 1991). Consider a LMM,

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{\epsilon}, \tag{2.1}$$

where $\boldsymbol{\epsilon}$ and $\boldsymbol{u}$ are mutually independent with $\boldsymbol{u} \sim N(\boldsymbol{0}, \boldsymbol{G}), \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \boldsymbol{R})$ with variance-covariance matrices $\boldsymbol{G}$(dimension: $h \times h$) and $\boldsymbol{R}(N \times N)$ is also known as conditional covariance matrix of $\boldsymbol{y}|\boldsymbol{X\beta} + \boldsymbol{Zu}$. In this model, the dimensions are $\boldsymbol{X}$ $(N \times p), \boldsymbol{\beta}(p \times 1), \boldsymbol{Z}$ $(N \times h), \boldsymbol{u}(h \times 1)$, and $\boldsymbol{\epsilon}(N \times 1)$ respectively, where $\boldsymbol{X}$ and $\boldsymbol{Z}$ are known design matrices, $\boldsymbol{y}$ $(N \times 1)$ is a vector of outcome measures. The response vector $\boldsymbol{y}$ is a linear combination of normally distributed random variables, hence the marginal distribution has the form of $\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \boldsymbol{ZGZ}^T + \boldsymbol{R})$, where $\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{\delta}) :=$

$\boldsymbol{ZGZ}^T + \boldsymbol{R}$, and $\boldsymbol{\delta} = \left(\delta_1, \dots, \delta_q\right)^T$ are variance parameters that covariance matrices $\boldsymbol{G}$ and $\boldsymbol{R}$ depend on.

GLMM considers both fixed and random effects to model continuous and discrete observations given that the observations are time independent. Unlike in general linear mixed models where $\boldsymbol{y}|\boldsymbol{u} \sim \boldsymbol{N}(\boldsymbol{X\beta} + \boldsymbol{Zu}, \boldsymbol{R})$, GLMM assumes that conditional response $\boldsymbol{y}|\boldsymbol{u}$ is not normally distributed, only $\boldsymbol{u} \sim N(\boldsymbol{0}, \boldsymbol{G})$**.** In such cases, estimation of marginal distribution of $\boldsymbol{y}$ is challenging and likelihood function does not have a closed form, leading to computationally heavy estimations. Existing methods to overcome this problem include the use of numerical methods to approximate the integrals of likelihood function through the Monte Carlo EM algorithms and Bayesian approach (Jiang & Lahiri, 2006; McCullagh, 2018).

For binary outcome data, the fixed and random effects in GLMM are linearly related with the mean of the outcome through logit link, while count data is related through log link. In recent years, GLMM is widely considered in different ways with small area estimations (Jiang & Lahiri, 2006; John NK Rao & Molina, 2015). Suppose that the variable of interest in small area estimation is a binary outcome, then the logit of probability of success for small areas i can be written as

$$\text{logit}\left(p_{ij}\right) = \text{logit}\left(\text{P}(y_{ij} = 1)\right) = \boldsymbol{X}_{ij}^T\boldsymbol{\beta} + u_i + e_{ij}, \tag{2.2}$$

where $i = 1, \dots, m, \quad j = 1, \dots, n_i$, $u_i \sim N(0, \sigma_u^2)$, $e_{ij} \sim N(0, \sigma_e^2)$, p_{ij} is the probability of $y_{ij} = 1$ for the element j in area i, u_i is the random effect for area i, e_{ij} is the residual of element j in area i. Direct calculation shows that

$$\frac{p_{ij}}{1 - p_{ij}} = \exp(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + u_i),$$

$$p_{ij} = \frac{\exp\left(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + u_i\right)}{1 + \exp(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + u_i)}.$$

2.1 BLUP and EBULP Estimators

The best linear unbiased prediction (BLUP) minimizes the mean squared error among the linear unbiased estimators. It does not depend on the normality of the random effects, but depend on the variance components. As stated in section 1.4, these BLUP parameters are estimated through method of moments (MOM), ML or REML. When this BLUP estimator is replaced by estimated variance components, it is referred to as empirical BLUP (EBLUP). Consider a linear combination in a form $\mu = \boldsymbol{l}^T\boldsymbol{\beta} + \boldsymbol{m}^T\boldsymbol{u}$, with regression parameters $\boldsymbol{\beta}$, random effects $\boldsymbol{u}$, and constant vectors $\boldsymbol{l}, \boldsymbol{m}$. A linear estimator $\hat{\mu}$ of $\mu = \boldsymbol{a}^T\boldsymbol{y} + b$ is unbiased if $E(\hat{\mu}) = E(\mu)$ (John NK Rao & Molina, 2015).

For given $\boldsymbol{\delta}$, the BLUP estimator of $\mu = \boldsymbol{l}^T\boldsymbol{\beta} + \boldsymbol{m}^T\boldsymbol{u}$ can be derived as (Henderson, 1950)

$$\tilde{\mu} = \boldsymbol{l}^T\tilde{\boldsymbol{\beta}} + \boldsymbol{m}^T\tilde{\boldsymbol{u}} = \boldsymbol{l}^T\tilde{\boldsymbol{\beta}} + \boldsymbol{m}^T\boldsymbol{G}\boldsymbol{Z}^T\boldsymbol{V}^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}\right), \tag{2.3}$$

where

$$\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}(\boldsymbol{\delta}) = (\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{y} \tag{2.4}$$

is the best linear unbiased estimator (BLUE) of $\boldsymbol{\beta}$, and

$$\tilde{\boldsymbol{u}} = \tilde{\boldsymbol{u}}(\boldsymbol{\delta}) = \boldsymbol{G}\boldsymbol{Z}^T\boldsymbol{V}^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}\right). \tag{2.5}$$

Assume that $\boldsymbol{u}$ and $\boldsymbol{\epsilon}$ follow multivariate normal distributions, the joint density of $\boldsymbol{y}$ and $\boldsymbol{u}$ is

$$\begin{aligned} f(\boldsymbol{y}, \boldsymbol{u}) &= f(\boldsymbol{y}|\,\boldsymbol{u}) \times \boldsymbol{f}(\,\boldsymbol{u}) \qquad (2.6)\\ &= (2\pi)^{-\frac{N}{2}}(\det(\boldsymbol{R}))^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u})^T\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u})\right] \\ &\quad \times (2\pi)^{-\frac{h}{2}}(\det(\boldsymbol{G}))^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\boldsymbol{u}^T\boldsymbol{G}^{-1}\boldsymbol{u}\right] \\ &\propto \exp\left[-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u})^T\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u}) - \frac{1}{2}\boldsymbol{u}^T\boldsymbol{G}^{-1}\boldsymbol{u}\right]. \end{aligned}$$

Assume the variance parameters $\boldsymbol{\delta} = \left(\delta_1, \ldots, \delta_q\right)^T$ are known, maximizing the joint likelihood of $\boldsymbol{y}$ and $\boldsymbol{u}$ with respect to $\boldsymbol{\beta}$ and $\boldsymbol{u}$ is equivalent to maximizing the joint log-likelihood

$$\ell(\boldsymbol{\beta},\boldsymbol{u}) = -\frac{1}{2}(\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{Zu})^T\boldsymbol{R}^{-1}(\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{Zu}) - \frac{1}{2}\boldsymbol{u}^T\boldsymbol{G}^{-1}\boldsymbol{u}. \quad (2.7)$$

Since $\boldsymbol{u}$ is unobservable, $\ell(\boldsymbol{\beta},\boldsymbol{u})$ can be considered as a penalized likelihood with a penalty term $\frac{1}{2}\boldsymbol{u}^T\boldsymbol{G}^{-1}\boldsymbol{u}$ added to the traditional log-likelihood function. The penalized log-likelihood function is conditioning on $\boldsymbol{u}$ as fixed (known).

In (2.2.5), by setting the partial derivative of $\ell(\boldsymbol{\beta},\boldsymbol{u})$ with respect to $\boldsymbol{\beta}$ and $\boldsymbol{u}$ to be zero, we have the following mixed model equations for both fixed effects and random effects

$$\begin{bmatrix}\boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{Z}\\ \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z}+\boldsymbol{G}^{-1}\end{bmatrix}\begin{bmatrix}\boldsymbol{\beta}^*\\ \boldsymbol{u}^*\end{bmatrix} = \begin{bmatrix}\boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{y}\\ \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{y}\end{bmatrix}. \quad (2.8)$$

The solution to (2.8) and BLUP estimators of $\boldsymbol{\beta}$ and $\boldsymbol{u}$ are identical, i.e. $\boldsymbol{\beta}^* = \widetilde{\boldsymbol{\beta}}$ and $\boldsymbol{u}^* = \widetilde{\boldsymbol{u}}$. Thus the BLUP estimators can also be considered as joint maximum likelihood estimators.

When the variance parameters $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_q)^T$ are unknown, the empirical BLUP (EBLUP) estimator of $\boldsymbol{u}$, $\widehat{\boldsymbol{u}} = t(\widehat{\boldsymbol{\delta}}, \boldsymbol{y}) = t(\widehat{\boldsymbol{\delta}})$ is obtained by replacing the unknown variance parameters $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_q)^T$ by an estimator $\widehat{\boldsymbol{\delta}} = \widehat{\boldsymbol{\delta}}(\boldsymbol{y})$.

2.2 ML and REML Estimators

Under the general linear mixed model, the maximum likelihood (ML) and restricted maximum likelihood (REML) estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ are obtained by maximizing the log likelihood function. Under the normality assumption, the log-likelihood is

$$\ell(\boldsymbol{\beta},\boldsymbol{\delta}) = -\frac{1}{2}(\boldsymbol{y}-\boldsymbol{X\beta})^T\boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{X\beta}) - \frac{1}{2}\log det\,(\boldsymbol{V}) + c, \quad (2.9)$$

where $c = -\frac{N}{2}\log(2\pi)$ is a constant, $\boldsymbol{V}$ is variance-covariance matrix of $\boldsymbol{y}\sim N(\boldsymbol{X\beta},\boldsymbol{V})$.

Take partial derivative of $\ell(\boldsymbol{\beta},\boldsymbol{\delta})$ with respect to $\boldsymbol{\beta}$, we have

$$\frac{\partial\ell(\boldsymbol{\beta},\boldsymbol{\delta})}{\partial\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{X\beta}) = \boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{y}-\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X\beta}.$$

Thus, MLE of $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{y}.$$

Take partial derivative of $\ell(\boldsymbol{\beta}, \boldsymbol{\delta})$ with respect to $\boldsymbol{\delta}$, we have

$$\frac{\partial \ell(\boldsymbol{\beta}, \boldsymbol{\delta})}{\partial \delta_j} = -\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T \frac{\partial \boldsymbol{V}^{-1}}{\partial \delta_j}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) - \frac{1}{2}\frac{\partial \log det(\boldsymbol{V})}{\partial \delta_j}$$

$$= \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{V}^{-1}\frac{\partial \boldsymbol{V}}{\partial \delta_j}\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) - \frac{1}{2}trace\left(\boldsymbol{V}^{-1}\frac{\partial \boldsymbol{V}}{\partial \delta_j}\right),$$

Note from Searle et al. (2009) (Searle, Casella, & McCulloch, 2009):

$$\frac{\partial \boldsymbol{V}^{-1}}{\partial \delta_j} = -\boldsymbol{V}^{-1}\frac{\partial \boldsymbol{V}}{\partial \delta_j}\boldsymbol{V}^{-1},$$

$$\frac{\partial |\boldsymbol{V}|}{\partial \delta_j} = trace\left(adj(\boldsymbol{V})\frac{\partial \boldsymbol{V}}{\partial \delta_j}\right),$$

$$\frac{\partial \log det\,(\boldsymbol{V})}{\partial \delta_j} = \text{trace}\left(\boldsymbol{V}^{-1}\frac{\partial \boldsymbol{V}}{\partial \delta_j}\right).$$

The expected second derivatives of $-\ell(\boldsymbol{\beta}, \boldsymbol{\delta})$ with respect to $\boldsymbol{\delta}$ is given by

$$\boldsymbol{I}_{jk}(\boldsymbol{\delta}) = E\left[-\frac{\partial^2 \ell(\boldsymbol{\beta}, \boldsymbol{\delta})}{\partial \delta_j \partial \delta_k}\right]$$

$$= E\left[-\frac{\partial\left(\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{V}^{-1}\frac{\partial \boldsymbol{V}}{\partial \delta_j}\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) - \frac{1}{2}trace\left(\boldsymbol{V}^{-1}\frac{\partial \boldsymbol{V}}{\partial \delta_j}\right)\right)}{\partial \delta_k}\right]$$

$$= E\left[-\frac{\partial\left(-\frac{1}{2}trace\left(\boldsymbol{V}^{-1}\frac{\partial \boldsymbol{V}}{\partial \delta_j}\right)\right)}{\partial \delta_k}\right]$$

$$= \frac{1}{2}trace\left(\boldsymbol{V}^{-1}\frac{\partial \boldsymbol{V}}{\partial \delta_j}\boldsymbol{V}^{-1}\frac{\partial \boldsymbol{V}}{\partial \delta_k}\right),$$

where $j = 1, \ldots, q$ and $k = 1, \ldots, q$ are the elements of the fisher Information matrix, $-\boldsymbol{I}_{jk}(\boldsymbol{\delta})$.

The MLEs $\widehat{\boldsymbol{\delta}}$ of $\boldsymbol{\delta}$ can be obtained iteratively by using Newton Raphson algorithm based on the first and second order partial derivatives of the log-likelihood function with respect to $\boldsymbol{\delta}$ and $\boldsymbol{\beta}$,

$$\boldsymbol{\delta}^{(n+1)} = \boldsymbol{\delta}^{(n)} + \left[\boldsymbol{I}\left(\boldsymbol{\delta}^{(n)}\right)\right]^{-1} \frac{\partial}{\partial \boldsymbol{\delta}} \ell(\boldsymbol{\beta}, \boldsymbol{\delta})|_{\widehat{\boldsymbol{\beta}}=\widetilde{\boldsymbol{\beta}}(\boldsymbol{\delta}^{(n)}), \widehat{\boldsymbol{\delta}}=\boldsymbol{\delta}^{(n)}},$$

where $n = 0,1, \ldots$, is the number of iterations.

The ML estimators of $\boldsymbol{\delta}$ and $\boldsymbol{\beta}$ at convergence are $\widehat{\boldsymbol{\delta}}$ and $\widehat{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\beta}}(\widehat{\boldsymbol{\delta}})$ respectively, where $\widehat{\boldsymbol{\delta}} = \boldsymbol{\delta}^{(n)}$and $\widehat{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\beta}}\left(\boldsymbol{\delta}^{(n)}\right)$ are values of $\boldsymbol{\delta}$ and $\widetilde{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\beta}}(\boldsymbol{\delta})$ at the n^{th} iteration. The asymptotic covariance matrix of ML estimators $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\delta}}$ is diag$\left[var\left(\widehat{\boldsymbol{\beta}}\right), var\left(\widehat{\boldsymbol{\delta}}\right)\right]$, where

$$\begin{aligned} var\left(\widehat{\boldsymbol{\beta}}\right) &= var[(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\,\boldsymbol{y}] \\ &= (\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{V}\,((\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1})^T \\ &= (\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{V}\,\boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1} \\ &= (\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}, \end{aligned}$$

and $var\left(\widehat{\boldsymbol{\delta}}\right) = \boldsymbol{I}^{-1}(\boldsymbol{\delta})$.

The major drawback of ML method is lack of considering loss of degrees of freedom due to one parameter when estimating for the other parameter. REML takes care of this issue by transforming data $\boldsymbol{y}^* = \boldsymbol{A}^T\boldsymbol{y}$, where $\boldsymbol{A}$ is any $N \times (N-p)$ full rank matrix orthogonal to $\boldsymbol{X}$. The REML estimators are obtained through restricted log-likelihood of the joint density of $\boldsymbol{y}^*$ expressed as a function of $\boldsymbol{\delta}$

$$\ell_R(\boldsymbol{\delta}) = -\frac{1}{2}\log \boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X} - \frac{1}{2}\boldsymbol{y}^T\boldsymbol{P}\boldsymbol{y} - \frac{1}{2}\log|\boldsymbol{V}| + c,$$

where $\boldsymbol{P} = \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}$.

REML estimators $\widehat{\boldsymbol{\delta}}_{RE}$ of $\boldsymbol{\delta}$ and $\widehat{\boldsymbol{\beta}}_{RE} = \widetilde{\boldsymbol{\beta}}(\widehat{\boldsymbol{\delta}}_{RE})$ of $\boldsymbol{\beta}$ also are obtained iteratively using Newton Raphson algorithm. The covariance matrices are asymptotically equal in both ML and REML estimators of $\boldsymbol{\delta}$ and $\boldsymbol{\beta}$ for fixed p, $var\left(\widehat{\boldsymbol{\delta}}\right) \approx var\left(\widehat{\boldsymbol{\delta}}_{RE}\right)$ and $var\left(\widehat{\boldsymbol{\beta}}\right) \approx var\left(\widehat{\boldsymbol{\beta}}_{RE}\right)$.

2.3 Empirical and Hierarchical Bayes Method for Small Area Models

Bayes techniques are based on basic Bayes theorem using marginal (prior distribution $f(\boldsymbol{\theta})$) and conditional probability density function (posterior distribution $f(\boldsymbol{y}|\boldsymbol{\theta})$). The marginal

likelihood of $\boldsymbol{y}$, $f(\boldsymbol{y}) = \int f(\boldsymbol{y}, \boldsymbol{\theta}) d\boldsymbol{\theta} = \int f(\boldsymbol{y}|\boldsymbol{\theta}) f(\boldsymbol{\theta}) d\boldsymbol{\theta}$ which involves integrals, becomes challenging with complicated posterior distributions and also with multivariate prior distributions. In such situations, parameters are obtained via simulation approaches, such as Markov Chain Monte Carlo (MCMC) approaches: Metropolis-Hastings, and Gibbs sampling. The Hierarchical Bayes (HB) method first simulate $\boldsymbol{\theta}$, then, simulate $\boldsymbol{y}$ given $\boldsymbol{\theta}$. The Empirical Bayes (EB) method is based on convenient prior distribution for $f(\boldsymbol{\theta}|\boldsymbol{\beta})$ known as conjugate prior, where $\boldsymbol{\beta}$ is a hyper parameter which is estimated by frequentist methods (Martuzzi & Elliott, 1996; John NK Rao & Molina, 2015; Yasui et al., 2000).

The Empirical Bayes approach in SAE assumes that the parameter of interest $\boldsymbol{\theta} = \boldsymbol{\mu}$, the population mean for the small area i in the linking model of FH model (equation (1.1)), has some prior distribution $f(\boldsymbol{\theta}|\boldsymbol{\beta})$ where $\boldsymbol{\beta}$ is an unknown parameter. First, the posterior distribution of $\boldsymbol{\theta}$, given data, is obtained assuming that $\boldsymbol{\beta}$ is known, then $\boldsymbol{\beta}$ is estimated using the marginal distribution of the data (Farrell, MacGibbon, & Tomberlin, 1997; Fay III & Herriot, 1979; Ghosh & Rao, 1994; Morris, 1983).

The Hierarchical Bayes model assumes that the parameter of interest $\boldsymbol{\theta}$ is from a prior distribution with some unknown parameters ($\boldsymbol{\beta}$) and again this unknown parameter has another prior distribution with unknown parameters. Therefore, in SAE, the basic area model under hierarchical Bayesian framework can be expressed as two stage hierarchical model which is also known as conditionally independent hierarchical model. Two models have discussed based on unknown sampling variance σ_e^2 and known sampling variance (unbiased estimate, s_e^2) in basic FH model (Y. You & Chapman, 2006). The authors showed that the results obtained from proposed HB method for the model with unknown variance performs well regardless of the sample size, comparing both models using two survey data sets, corn-soybean and milk data from U.S. Department of Agriculture and U.S. Bureau of Labor Statistics respectively.

The basic FH model (equation (1.1)) assumes that the direct estimate y_i and sampling variance σ_e^2 of $e_i \sim N(0, \sigma_e^2)$ are known and obtained from auxiliary data. This assumption could be too strong and lead to biased results especially for the areas with small sample sizes. You and Chapman (2006) introduced a Hierarchical Bayes (HB) model using Gibbs sampling technique to estimate unknown σ_e^2 from an unbiased estimator s_e^2 assuming that the s_e^2 is independent of the direct estimator y_i, and $(n_i - 1)s_e^2 \sim \sigma_e^2 \chi_{n_i-1}^2$, where n_i is the sample size for area i (Y. You & Chapman, 2006). Some applications of empirical and hierarchical Bayes methods in SAE are discussed in detail under the literature review in the section 2.6 (Ghosh, Natarajan, Stroud, & Carlin, 1998; Hobza & Morales, 2016).

2.4 Hierarchical Generalized Linear Models for SAE Models

2.4.1 Hierarchical Generalized Linear Models

A hierarchical generalized linear model (HGLMs) uses a generalization of Henderson's joint likelihood to allow components of random effects in the linear predictors of generalized linear models. For example, the distribution of random effect u_i for area i in Poisson-Gamma HGLM (or Poisson conjugate HGLM) is Poisson with mean λ, and the distribution of $\boldsymbol{y}|u_i$ is gamma, which is linked through a canonical log function . Some other examples of HGLMs are Binomial-Beta, Gamma-Inverse Gamma, and Inverse Gaussian-Gamma HGLMs. Normal-Normal or Normal conjugate HGLM is a special case of HGLM with identity link function, where u_i is normally distributed, also known as LMM (Lee, Nelder, & Pawitan, 2006; Lee & Nelder, 1996).

Lee and Nelder (1996) originally defined HGLM as a GLM family for the response variable $\boldsymbol{y}$ given a random effect $\boldsymbol{u}$, satisfying

$$E(\boldsymbol{y}|\boldsymbol{u}) = \mu \text{ and } var(\boldsymbol{y}|\boldsymbol{u}) = \phi V(\mu),$$

with conditional log likelihood of $\boldsymbol{y}$ given $\boldsymbol{u}$

$$\ell(\theta, \phi; \boldsymbol{y}|\boldsymbol{u}) = \frac{\sum\{\boldsymbol{y}\theta - b(\theta)\}}{\phi} + c(\boldsymbol{y}, \phi), \tag{2.10}$$

where $\theta = \theta(\mu)$ is the canonical (natural) parameter, μ is the conditional mean of $\boldsymbol{y}$ given $\boldsymbol{u}$, ϕ is the dispersion parameter. The linear predictor takes the form of

$$\eta = g(\mu) = \boldsymbol{X\beta} + \boldsymbol{Zv},$$

where $g()$ is the link function, $\boldsymbol{\beta}$s are the regression coefficients of fixed effects and random component $\boldsymbol{v} = v(\boldsymbol{u})$ is a monotonic function of random effects $\boldsymbol{u}$. The random effects $\boldsymbol{u}$ is extended to conjugate distribution from the GLM family with parameters α. This will be a key advantage of HGLMs compared to GLMs.

HGLM approach has been widely used in modelling binary and count data, frailty modelling for survival data, repeated measures data, and survey data in both univariate, multivariate cases (Ha, Lee, & Song, 2001; Molenberghs, Verbeke, Demétrio, & Vieira, 2010).

Normal-Normal HGLM

Normal-Normal HGLM is a GLM with $\boldsymbol{y}|\boldsymbol{u} \sim N(\boldsymbol{X\beta} + \boldsymbol{Zv}, \sigma^2), \boldsymbol{u} \sim N(0, \lambda), \lambda = \sigma_u^2$ and the identity link function $(g(\mu) = \mu)$, hence $\boldsymbol{v} = \boldsymbol{u}$, and the log likelihood of $\boldsymbol{y}|\boldsymbol{u}$

$$\ell(\theta, \phi; \boldsymbol{y}|\boldsymbol{u}) = -\frac{1}{2\sigma^2}\big(\boldsymbol{y} - (\boldsymbol{X\beta} + \boldsymbol{Zv})\big)^2 - \log 2\pi\sigma^2$$

$$= \left\{\frac{\boldsymbol{y}(\boldsymbol{X\beta} + \boldsymbol{Zv}) - (\boldsymbol{X\beta} + \boldsymbol{Zv})^2/2}{\sigma^2} - \left(\frac{\boldsymbol{y}}{2\sigma^2} + \log 2\pi\sigma^2\right)\right\},$$

has the form of (2.10), where $var(\boldsymbol{y}|\boldsymbol{u}) = \phi = \sigma^2, \theta = \boldsymbol{X\beta} + \boldsymbol{Zv}, b(\theta) = (\boldsymbol{X\beta} + \boldsymbol{Zv})^2, V(\mu) = 1, \eta = g(\mu) = \boldsymbol{X\beta} + \boldsymbol{Zv}, \boldsymbol{v} = \boldsymbol{u}, \boldsymbol{u} \sim N(0, \lambda), \lambda = \sigma_u^2,$ and $c(\boldsymbol{y}, \phi) = -(\boldsymbol{y}/2\sigma^2 + \log 2\pi\sigma^2)$. Here, the random component $\boldsymbol{v}$ has the normal distribution with identify link function.

Poisson-Gamma HGLM

Poisson-Gamma HGLM is an extension of GLM with the random effect $(\boldsymbol{u})$ has a Gamma distribution, and $\boldsymbol{y}|\boldsymbol{u} \sim Poisson(\boldsymbol{X\beta} + \boldsymbol{v})$. The log likelihood of $\boldsymbol{y}|\boldsymbol{u}$

$$\ell(\theta, \phi; \boldsymbol{y}|\boldsymbol{v}) = \boldsymbol{y}\log(\boldsymbol{X\beta} + \boldsymbol{v}) - (\boldsymbol{X\beta} + \boldsymbol{v}) - \log \boldsymbol{y}!$$

where $\theta = \log(\boldsymbol{X\beta} + \boldsymbol{v})$, $b(\theta) = \boldsymbol{X\beta} + \boldsymbol{v}$, $\phi = 1, b(\theta) = \exp\theta = \boldsymbol{X\beta} + \boldsymbol{v}$, and $\boldsymbol{v} = \log \boldsymbol{u}$. The random component $\boldsymbol{v}$ has the log-gamma distribution with link function being the log link and $\boldsymbol{u}$ being gamma distribution in Poisson-Gamma HGLM. The expected value of $\boldsymbol{y}|\boldsymbol{u}, E(\boldsymbol{y}|\boldsymbol{u}) = b'(\theta) = \boldsymbol{X\beta} + \log \boldsymbol{u}$.

2.4.2 H-likelihood and MHLE of $(\boldsymbol{\beta}, \boldsymbol{u})$

H-likelihood is the logarithm of the joint density function of $\boldsymbol{y}$ and $\boldsymbol{v}\,(= v(\boldsymbol{u}))$, equivalently, it is the joint density function of $\boldsymbol{y}$ and $\boldsymbol{u}$, which needs to be the form of

$$h = \log f_{\beta,\phi}(\boldsymbol{y}|\boldsymbol{u}) + \log f_{\alpha}(\boldsymbol{u}) = \ell(\theta, \phi; \boldsymbol{y}|\boldsymbol{u}) + \ell(\alpha|\boldsymbol{u}),$$

where $\ell(\theta, \phi; \boldsymbol{y}|\boldsymbol{u})$ is the logarithm of the density function of $\boldsymbol{y}|\boldsymbol{u}$, $\ell(\alpha|\boldsymbol{u})$ is that of $\boldsymbol{u}$, $\boldsymbol{\beta}$ are fixed effects, ϕ are dispersion parameters, and α are parameters for random effects. The joint likelihood is used to estimate BLUP of random effects $\boldsymbol{u}$ and MLE of fixed effects $\boldsymbol{\beta}$, is straightforward in standard linear mixed model with $\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$ and $\boldsymbol{u} \sim N(0, \sigma_u^2)$. When the joint log likelihood does not have a closed form, it is very challenging to estimate BLUP or EBLUP. In such situations h-likelihood comes into play. The hierarchical(h) maximum likelihood estimates (HMLEs) are obtained by score equations $\partial h/\partial \boldsymbol{\beta} = 0$, and $\partial h/\partial \boldsymbol{u} = 0$. Unlike in GLM, the random component $\boldsymbol{v}$, or equivalently, the random effect $\boldsymbol{u}$ in HGLM is not assumed to be normally distributed, instead, it is estimated by the properties of data through a prior distribution. However, the HMLEs are often obtained via numerical approximation methods due to intractable integrals in the log-likelihood function.

The inferences about $\boldsymbol{\beta}, \boldsymbol{u}$, and the dispersion parameters (ϕ) in HGLMs involve three likelihoods; h-likelihood, and two adjusted profile likelihoods (marginal likelihood and restricted likelihood), where inference about $\boldsymbol{\beta}$ is based on marginal likelihood $L = \log \int \exp(h) d\boldsymbol{u}$, $\boldsymbol{u}$ is based on h-likelihood and the dispersion parameters are based on restricted likelihood, respectively.

Let $\boldsymbol{\tau} = (\boldsymbol{\beta}, \boldsymbol{u})$ be the fixed effects parameters. Given the variance parameter $\theta = \boldsymbol{\delta} = (\delta_1, \dots, \delta_q)^T$, the maximum hierarchical (h)-likelihood estimators (MHLE) for $\hat{\boldsymbol{\tau}} = (\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{u}})$ are obtained by solving the score function $\partial h/\partial \boldsymbol{\tau} = 0$. If the solution does not have a closed form, we can use Newton-Raphson methods to generate an iterative procedure that uses the gradient vector and the observed information matrix to approximate the points that maximize a likelihood function. Start at initial values $(\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}^{(0)}, \boldsymbol{u} = \widehat{\boldsymbol{u}}^{(0)})$, the approximate maximums are updated iteratively using

$$\begin{pmatrix} \widehat{\boldsymbol{\beta}}^{(k+1)} \\ \widehat{\boldsymbol{u}}^{(k+1)} \end{pmatrix} = \begin{pmatrix} \widehat{\boldsymbol{\beta}}^{(k)} \\ \widehat{\boldsymbol{u}}^{(k)} \end{pmatrix} + \left(\boldsymbol{\mathcal{J}}^{-1}\boldsymbol{\mathcal{S}}(\tau)\right)|_{(\boldsymbol{\beta},\boldsymbol{u})=(\widehat{\boldsymbol{\beta}}^{(k)},\widehat{\boldsymbol{u}}^{(k)})}, \tag{2.11}$$

where $\boldsymbol{\tau} = (\boldsymbol{\beta}, \boldsymbol{u})$, $\boldsymbol{\mathcal{S}}(\boldsymbol{\tau}) = \begin{pmatrix} \frac{\partial h}{\partial \boldsymbol{\beta}} \\ \frac{\partial h}{\partial \boldsymbol{u}} \end{pmatrix}$ is the score function, $\boldsymbol{\mathcal{J}} = \begin{pmatrix} -\left[\frac{\partial^2 h}{\partial^2 \boldsymbol{\beta}}\right]_{p\times p} & -\left[\frac{\partial^2 h}{\partial \boldsymbol{\beta}\partial \boldsymbol{u}}\right]_{p\times m} \\ -\left[\frac{\partial^2 h}{\partial \boldsymbol{u}\partial \boldsymbol{\beta}}\right]_{p\times p} & -\left[\frac{\partial^2 h}{\partial^2 u}\right]_{m\times m} \end{pmatrix}$ is the asymptotic covariance matrix of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{u}}$. The variance-covariance matrices will be estimated by MHLE of $\boldsymbol{\theta}$ of adjusted h-likelihood through an iterative procedure which is described in detail in chapter 4.

2.4.3 Penalized Partial Maximum Likelihood Estimation of dispersion parameters

In general, the variance components and dispersion parameters are estimated from the REML approach by maximizing the REML likelihood $f_{\boldsymbol{\theta}}(\boldsymbol{y}|\widehat{\boldsymbol{\beta}})$ which is straightforward in linear mixed model. For most cases, marginal and joint likelihoods do not have a closed form, hence it requires some approximation to compute the likelihoods. In such situations, the likelihood function is approximated via integral approximation methods, such as EM algorithm, Gibbs sampling, MCMC type algorithms, Laplace approximation, and h-likelihood approximation through Laplace approximation(Laplace, 1986; Lee et al., 2006). Some researchers have shown that Laplace approximation is computationally efficient and less bias compared with Gibbs sampling and MCMC algorithms with high dimensional integration approximations (Breslow & Clayton, 1993;

Lee & Nelder, 1996; Noh & Lee, 2007). Lee and Nelder (2006) introduced an alternative approach to obtain dispersion parameters using h-likelihood through Laplace approximation (Lee et al., 2006).

The Laplace approximation is used to approximate an integral of the form $\int_{x_0}^{x_1} e^{cf(x)}dx$, where $f(x)$ is a twice differentiable function, and c is a constant. Suppose that $f(x)$ is a continuous, differentiable function, $\lim_{x\to x_0} f(x) \to f(x_0)$, and $f''(x_0) < 0$. From, Taylor series expansion,

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2}f''(x_0)(x - x_0)^2 + \mathcal{O}((x - x_0)^3),$$

which can be expressed as

$$f(x) \approx f(x_0) - \frac{1}{2}|f''(x_0)|(x - x_0)^2, \tag{2.12}$$

where $\lim_{x\to x_0} f'(x) \to 0$. The equation (2.12) can be simplified as

$$\int_{x_0}^{x_1} e^{f(x)}\,dx \approx e^{f(x_0)} \int_{x_0}^{x_1} e^{-\frac{1}{2}|f''(x_0)|(x-x_0)^2}\,dx.$$

The above integral has the form of an arbitrary Gaussian integral function. Since $|f''(x_0)| > 0$,

$$\int_{\mathbb{R}^n} e^{-\frac{1}{2}n|f''(x_0)|(x-x_0)^2}dx = \lim_{n\to\infty}\left(\frac{2\pi}{n|f''(x_0)|}\right)^{1/2}.$$

Now, the Laplace approximation to $\int_{x_0}^{x_1} e^{nf(x)}\,dx$ can be written as

$$\int_{x_0}^{x_1} e^{nf(x)}\,dx \approx e^{nf(x_0)} \lim_{n\to\infty}\left(\frac{2\pi}{n|f''(x_0)|}\right)^{1/2}$$

which can be shown using the lower and upper bounds from Taylor's theorem as

$$\lim_{n\to\infty} \frac{\int_{x_0}^{x_1} e^{nf(x)}\,dx}{e^{nf(x_0)}\left(\frac{2\pi}{n(-f''(x_0))}\right)^{\frac{1}{2}}} = 1$$

where $f(x)$ has a global maximum at $x = x_0$, then $f''(x_0) < 0$. This expression can be easily generalized to first-order Laplace approximation as

$$\int e^{f(x)}dx \approx e^{f(x_0)}\left\{\left|-\frac{1}{2\pi}f''(x)\right|^{-\frac{1}{2}}\right\}\Bigg|_{x=x_0},$$

where x_0 is a global maximum of some function $f(x)$. This technique is used defining an adjusted h-likelihood h_A where it is cumbersome to approximate the integrals when obtaining the REML and marginal likelihoods.

Take the log transformation of Laplace approximation

$$\log\int \exp f(x)\,dx \approx \log\left\{\left|-\frac{1}{2\pi}\frac{\partial^2 f(x)}{\partial x^2}\right|^{-\frac{1}{2}}\exp f(x)\right\}\Bigg|_{x=x_0}$$

$$= f(x)|_{x=x_0} - \frac{1}{2}\left\{\log\det\left(\frac{\mathcal{J}}{2\pi}\right)\right\}\Big|_{x=x_0}$$

$$\int f(x)\,dx = f(x)|_{x=x_0} + \frac{1}{2}\{\log\det(2\pi\mathcal{J}^{-1})\}|_{x=x_0},$$

where $\mathcal{J} = \partial^2 f(x)/\partial x^2$. This concept was used proposing the adjusted h-likelihood h_A to approximate the REML log likelihood $f_{\boldsymbol{\theta}}(\boldsymbol{y}|\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{u}})$ (Lee et al., 2006)

$$h_A = h|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}},\boldsymbol{u}=\widehat{\boldsymbol{u}}} + \frac{1}{2}\log\{\det(2\pi\boldsymbol{\mathcal{J}}^{-1})\}|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}},\boldsymbol{u}=\widehat{\boldsymbol{u}}}$$

$$= h|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}},\boldsymbol{u}=\widehat{\boldsymbol{u}}} + \frac{1}{2}\log\{(2\pi)^{p+m}\det(\boldsymbol{\mathcal{J}}^{-1})\}\Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}},\boldsymbol{u}=\widehat{\boldsymbol{u}}}$$

$$= h|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}},\boldsymbol{u}=\widehat{\boldsymbol{u}}} + \frac{1}{2}(p+m)\log 2\pi - \frac{1}{2}\log\det(\boldsymbol{\mathcal{J}})\Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}},\boldsymbol{u}=\widehat{\boldsymbol{u}}}, \qquad (2.13)$$

which is also known as the penalized partial likelihood function (PPL). Now, consider the first order of Laplace approximation of PPL and solve $\partial h/\partial u = 0$ which results the MLE for θ_r. By maximizing adjusted profile h-likelihood with respect to $\boldsymbol{\theta}$. Since $\boldsymbol{\mathcal{J}}$ is invertible, we have the score function $\partial h_A/\partial \theta_r$ as

$$\frac{\partial h_A}{\partial \theta_r} = \frac{\partial h|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}},\boldsymbol{u}=\widehat{\boldsymbol{u}}}}{\partial \theta_r} - \frac{\partial}{\partial \theta_r}\left(\frac{1}{2}\log\{\det(\boldsymbol{\mathcal{J}})\}|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}},\boldsymbol{u}=\widehat{\boldsymbol{u}}}\right)$$

$$= \frac{\partial h|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}},\boldsymbol{u}=\widehat{\boldsymbol{u}}}}{\partial \theta_r} - \frac{1}{2} trace\left(\boldsymbol{\mathcal{J}}^{-1}\frac{\partial \boldsymbol{\mathcal{J}}}{\partial \theta_r}\right)\Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}},\boldsymbol{u}=\widehat{\boldsymbol{u}}}. \quad (2.14)$$

The MHLE of θ_r is obtained solving $\partial h_A/\partial \theta_r = 0$ iteratively updating $\widehat{\boldsymbol{\beta}}$, and $\widehat{\boldsymbol{u}}$ which are obtained through Newton Raphson method.

2.5 Literature Review on Advanced Modelling Techniques in Small Area Estimation

Many researchers have adopted Hierarchical Generalized Linear modelling (HGLM) approach in the past to make accurate inferences in SAE since its' flexibility of analyzing continuous and discrete data. In recent years, the Empirical Bayes (EB) and Hierarchical Bayes (HB) related GLM methods are also proposed and extensively discussed the robustness of these estimators against that of different types of SAE estimation methods. The EB approach assumes that the population data are a sample from a super population which can be represented by an empirical Bayes model. This assumption clearly states the issues of repeated survey analysis where the population of inference might be different across the time period of survey conducted. Such situations have been handled via time series methodology in SAE expanding the empirical Bayes method borrowing information from previous surveys as well as from other areas(Jon NK Rao & Yu, 1994).

While many authors have conducted research studies focusing on cross sectional data, it has been increased the interest of repeated measure studies as well. Rao and Yu (1994) proposed an extension of basic area level FH model considering auto correlated random effects and sampling

error for SAE using time series and cross-sectional data. The two-stage estimator of small area mean for the current period is obtained under two scenarios; I. for known autocorrelation, the BLUP is first obtained and then replaced with their consistent estimators, 2. for unknown autocorrelation, three methods proposed to estimate autocorrelation via simulation approach (Jon NK Rao & Yu, 1994).

Hierarchical Bayes (HB) methodology in SAE has been widely considered in different scenarios (Ghosh et al., 1998; Stroud, 1994). Ghosh et al. (1998) provided a theorem covering binary, count, multi category, and spatial data based on hierarchical Bayes GLMM and implemented via MCMC approximation method, in particular through the Gibbs sampling technique using two real datasets. One dataset with multi-category outcome variable representing any negative impact of experiencing health hazards exposure in the workplace (yes, no, not exposed, and not applicable or not stated) is considered along with other demographic information, age, sex, region in 15 regions in Canada on a 1991 sample. The authors compared the estimates from the simulation approach with the estimates from the multi-category logistic regression model, describing the standard errors of the HB estimates are much lower than the sample proportions. The second data set relates to cancer mortality rates due to lung cancer in 115 counties in Missouri during 1972-1981 (Ghosh et al., 1998).

Many researchers have proposed empirical Bayes SAE approaches with the assumption of population data is a sample from the larger super population (Dempster & Tomberlin, 1980; Fay III & Herriot, 1979; MacGibbon & Tomberlin, 1987; Morris, 1983; Yasui et al., 2000). The EB methodology for estimation of small area proportions or binary responses is first proposed by Dempster et. al (1980). The technique was illustrated incorporating random effects and nested random effects to estimate census undercount of small groups of population. The U.S. Bureau of Census has conducted matching studies in each census since 1950, collecting data from external sources such as administrative records close to the census date. It is then compared with appropriate census group and the missing proportion in census records is considered as undercount for relevant

subgroups. The probability of an individual being in a census is estimated as a function of demographic characteristics such as age, sex, and race categories using logistic model introducing a local area effect. The disadvantage of this model is being able to include data only within local area and can only obtain estimates effects and interaction for local areas. The authors extended the model considering the prior distributions for random effects of primary sampling units (PSU), secondary sampling units (SSU) within PSUs, and tertiary sampling units within SSUs are assumed to be normal with zero mean and different variances. The corresponding EB logistic model is adopted to estimate census undercount from an approximate posterior distribution of undercount (Dempster & Tomberlin, 1980).

Furthermore, MacGibbon et.al. (1989) considered empirical Bayes SAE technique proposed by Dempster et. Al (1980) to estimate small area proportions including random effects and nested random effects. The EB estimates are obtained through a MC simulation study and compared with the estimated proportions based on classical unbiased estimates, model-based estimates (synthetic estimates), and empirical Bayes estimates using a labor force participation data from 15 PSUs. The auxiliary data were generated with identical distributions, uniformly distributed age, and Bernoulli distributed sex variable with proportion 0.5 for each group. The model based or the synthetic estimate was obtained based on fixed effects logistic model considering only local area level data. However, the synthetic estimators are biased due to lack of capturing the between area effects. This is solved from the empirical Bayes logistic model where random effects are assumed to have a multivariate normal distribution with mean zero and the variance σ_u^2. The model first assumed the variance component of the random effects σ_u^2, then estimate the proportions or the posterior distribution of random effects from the simulated data set. It is then used to obtain the maximum likelihood function of σ_u^2, and estimate MLE $\widehat{\sigma_u^2}$ via EM algorithm which is used to estimate the posterior variances (MacGibbon & Tomberlin, 1987).

The binomial logit modelling in SAE has been considered through various estimation approaches to estimate fixed, random effects estimates using maximum likelihood estimation methods, Bayesian inferential methods, and methods of moments (MM) etc. Hobza and Morales (2016) adopted unit-level binomial logit mixed model to estimate empirical best prediction of weighted sums of probabilities via MCMC simulation (Hobza & Morales, 2016). The random effects is assumed to be independent and identically distributed $\boldsymbol{u} = \left(u_{1,\dots,}u_m\right)^T \sim N(\mathbf{0}, \boldsymbol{I}_m)$ where $\boldsymbol{I}_m$ unit matrix is the variance covariance matrix of random effects, and $y_{ij}|u_i \sim Bin\left(n_{ij}, p_{ij}\right), i = 1 \dots, m, j = 1, \dots, n_i$. The variance parameter and the regression parameters are estimated using the method of simulated moments (MSM) through Newton-Raphson iterative formula. Furthermore, the empirical best predictor for p_{ij} and the weighted sum of probabilities for small area i were approximated via Monte Carlo (MC) simulation. Last, the mean squares error (MSE) and the error corrections of the parameters were obtained via MM and MC simulation.

The hierarchical likelihood approach is not only widely used in estimating the random effects and fixed effects, but also used in error correction which occurs in the process of estimating random effects and fixed effects. Lee et al. (2011) proposed hierarchical likelihood prediction intervals for random effects and fixed effects using h-likelihood approach. (Lee, Jang, & Lee, 2011) The authors showed that the prediction interval from HL is very accurate compared to the results from penalized quasi likelihood, and fully Bayesian methods using a lip cancer dataset in areas of Scotland and an infant mortality dataset in British Columbia, Canada.

Furthermore, HGLMs are also considered in Bayesian approaches through MCMC methods to estimate parameters, the methodology is illustrated using state-level and hospital-level auxiliary data to describe the cluster specific rates of utilization for both hospitals and states. Huang and Wolfe considered h-likelihood using EM algorithm together with MCMC, and stated that it can also be considered as a modification of the MCMC. These approaches discuss the variations due to clustering and the cluster size. Additionally the estimates can be used to draw conclusions

on higher levels using the hierarchical structure of the data (Daniels & Gatsonis, 1999; Huang & Wolfe, 2002). Ghosh et al. (1998) considered Hierarchical Bayes GLM to model discrete and continuous data using MCMC integration to obtain the joint posterior distribution avoiding high dimensional numerical integration (Ghosh et al., 1998).

Most of the researchers have considered uncorrelated random effects, but it is reasonable to consider the correlation between neighboring areas in many practical application problems. The correlation approaches to zero when the distance between neighborhood areas increases. Spatial hierarchical models are considered when the random effects $u_i, i = 1, \dots, m$ are not iid, i.e. they are correlated. The most common spatial small area model is conditional auto-regression (CAR) spatial model which assumes that the conditional distribution of small area i $Z_i u_i$ in equation (1.1), given the area effects for the other areas (neighboring areas for area i) can be obtained using the information of area i, where $Z_i = 1/\sqrt{C_i}$ in CAR spatial model, C_i is census count for small area i, and $Z_i = 1$ in basic FH model. The CAR spatial model is used in estimating US Census undercount of certain areas based on spatial dependence (Cressie, 1991).

In 2006, Alessandra et al. considered spatially correlated random effects model and proposed a methodology based on EBLUP estimator using a simultaneously autoregressive (SAR) model (Petrucci & Salvati, 2006). The authors proposed an estimator for MSE combining EBLUP with a SAR model, called as spatial EBLUP ($\theta(\hat{\sigma}_u^2, \hat{\rho})$) estimators which were evaluated via MC simulation of spatially correlated random effects data using a soil erosion data set in south-west watershed. Considering the estimated spatial autocorrelation coefficients from ML and REML method shows the existence of a strong spatial relationship between random effects. Under SAR model, synthetic estimator in equation (1.1) has the form $\widehat{\boldsymbol{\theta}} = \boldsymbol{X\beta} + \boldsymbol{Z}(\boldsymbol{I} - \rho\boldsymbol{W})^{-1}\boldsymbol{u} + \boldsymbol{\epsilon}$, where $\boldsymbol{W}$ is the spatial weight matrix for $\boldsymbol{\theta}$, $\rho \in (-1,1)$ is the spatial autoregressive coefficient.

Most importantly, it is required to concentrate on point estimates as well as prediction errors for each small area, but not the average prediction error in SAE problems. The prediction

error or the accuracy of the estimates plays a big contribution to the accuracy of estimates since SAE considers important practical applications, such as health disease estimation, fund allocation estimation, environmental health related estimations, US census undercount estimation, etc. Many researchers have proposed prediction MSE (PMSE) estimators using various estimation techniques (Jiang, Lahiri, & Wan, 2002; N. Prasad & Rao, 1986; N. N. Prasad & Rao, 1990). In 1990, Prasad et al. developed PMSE estimators which occur when estimating EBLUP of $\boldsymbol{\beta}$ and variance components $\boldsymbol{\theta}$ under the linear mixed models where the random effects are independent normally distributed with variance σ_u^2 which is estimated using MOM method (N. N. Prasad & Rao, 1990).

Chapter 3. Multilevel Small Area Estimation in Survey Research

Small area estimation is widely applied survey research. Small-area estimation (SAE) is a statistical technique used to produce statistically reliable estimates for smaller geographic areas (e.g., counties or sub-county areas) than those for which the original surveys were designed (Rahman & Harding, 2016; John NK Rao & Molina, 2015). Existing studies of SAEs that were developed using adult data from Behavioral Risk Factor Surveillance System (BRFSS) have been focusing on estimation of adult smoking prevalence.(Dwyer-Lindgren et al., 2014) In this chapter, we will apply the method to estimate youth electronic cigarette use prevalence at the county level using the combined 2014 and 2015 National Youth Tobacco Survey (NYTS).

We will perform a multilevel SAE model to incorporate individual-level tobacco use behaviors and area-level (county and state) ecological characteristics to electronic use prevalence among youth at the community level (i.e., county). Our SAE model is based on the analytical framework in conjunction with the mixed modeling of random effects and post-stratification simulation at the county level. This analytical design is considered superior to individual only or ecological only studies, which could yield mixed results as they miss the other important component in the analysis (Wills & Soneji, 2018).

3.1 Emerging tobacco use among adolescents

Tobacco use is the leading cause of preventable death (Centers for Disease Control and Prevention, 2013; U.S. Department of Health and Human Services, 2014) with the vast majority of tobacco use beginning during adolescence. In 2017, an estimated 3 million middle and high school students were current tobacco product users (T. W. Wang et al., 2018). In the United States, 9 out of 10 current smokers started smoking before 18 (U.S. Department of Health and Human Services, 2012). Nicotine exposure during adolescence can harm brain development and tobacco use at an

early age is associated with lower rates of smoking cessation and increased risk of addiction and other substance use including marijuana (Taioli & Wynder, 1991; U.S. Department of Health and Human Services, 2012, 2014).

The tobacco use landscape of youth has substantially changed in recent years with more adolescents using e-cigarettes and other emerging tobacco products (Jamal et al., 2017). Among U.S. high school students, current use of electronic cigarettes (e-cigarettes) has outpaced the use of traditional cigarettes (Jamal et al., 2017; U.S. Department of Health and Human Services, 2016). E-cigarettes contain varying levels of nicotine and a number of potentially toxic substances (Cameron et al., 2014; Dinakar & O'Connor, 2016; Goniewicz, Hajek, & McRobbie, 2014; Jensen, Luo, Pankow, Strongin, & Peyton, 2015) and youth use of e-cigarettes might serve as a gateway for future cigarette use (Barrington-Trimis et al., 2016; Leventhal et al., 2015; Primack, Soneji, Stoolmiller, Fine, & Sargent, 2015; Soneji et al., 2017; Wills et al., 2016). Cigars currently ranked third among the most commonly used tobacco products (Jamal et al., 2017) and cigar smoke is possibly more toxic than cigarette smoke (Institute, 1998; U.S. Department of Health and Human Services, 2012). Flavorings are frequently added to cigars to enhance their appeal to youth (Kostygina, Glantz, & Ling, 2016).

A growing body of literature has evaluated youth emerging tobacco use at the national level and identified patterns and socio-economic factors associated with youth substance use (H. Dai, 2018; H. Dai, 2019; H. Dai, D. Catley, K. P. Richter, K. Goggin, & E. F. Ellerbeck, 2018). However, no study has evaluated prevalence of emerging tobacco use among adolescents. Although enormous progress has been made in reducing tobacco use in the US, this progress has not been equally distributed across population with large disparities in tobacco use persisting across groups defined by race/ethnicity, education level, income level, region, and other factors (National Cancer Institute, 2017; N. A. Rigotti & S. Kalkhoran, 2017). Thus, assessing tobacco related health disparities using small area modelling can provide information for policy makers and stakeholders to develop interventions in curbing emerging tobacco use among adolescents.

3.2 Materials and Methods

We used data on National Youth Tobacco Survey (NYTS) 2014-2015 tobacco use in 143 counties downloaded from Centers for Disease Control and Prevention (CDC) and US Census 2015. NYTS is a FDA approved science based survey approach to study public health issues with tobacco use which includes data on long term, intermediate and short term relevant to tobacco usage. Even though cigarettes were introduced to reduce the tobacco use as well as to decrease the health issues, it has been addressed that the increase of electronic cigarette usage among youth due to many reasons, such as flavors of cigarettes, being less harmful than other tobacco, etc. Our main focus is on the prevalence of two variables depending on current and ever use of electronic cigarettes, 1. ever use of electronic cigarettes (yes or no), 2. current use of electronic cigarettes (yes or no). The original NYTS data had 39718 unit level observations from 143 counties in 41 states in the United States.

We extracted poverty data for 3220 counties in 53 states from American Community Survey (ACS) for the period 2011-2015 US Census. We considered 4 demographic groups for race, white, black, Hispanic races itself and others as the combination of "Asian alone", "American Indian and Alaska Native alone", and "Native Hawaiian and other Pacific Islander alone". County wise sub population for age x sex x race cross-tabulated data for 16 demographic categories were extracted from 2015 5-year API Census data using tidycensus R package (Walker, 2018). The fully adjusted sampling weight is taken adjusting the sampling weight which is the inverse of the probability of selection to alleviate the effect from non-respondents.

3.2.1 GLMM Model Specification

We consider the GLMM to obtain EBLUPs for county-wise prevalence as follows

$$\text{logit}\big(P_{ijkc}\big) = \text{logit}\big(\text{P}(y_{ijkc} = 1|u)\big) = \alpha_i + \beta_j + \gamma_k + x_c'\eta + \mu_c + e_{ijkc}, \qquad (3.2.1)$$

where α_i, β_j, and γ_k are regression coefficients of age group $i = 1,2$ ($1 = 10-14, 2 = 15-19$ years), sex $j = 1,2$ ($1 = \text{male}, 2 = \text{female}$), and race $k = 1, \dots, 4$ ($1 = \text{white}, 2 = \text{black}, 3 = \text{hispanic}, 4 = \text{others}$) respectively. x_c is county level covariates and η is corresponding regression coefficients for county level covariates. μ_c is county level random effects, and e_{ijkc} is residual. We only consider the county level poverty rate for x_c and obtained two models to obtain EBLUPs for E-cigarette current usage and E-cigarette ever usage using FH method and HB approach. The poverty rates were calculated using 2011-2015 ACS data extracted from US census(ACS)(ACS) (ACS). First, we started with sae and BayesSAE R packages to obtain county level estimates, but we could not use RStudio due to unavailability of assigning reference group for discrete auxiliary data in SAE estimation (Molina & Marhuenda, 2015; Shi & Shi, 2013). We obtained county level random effects and regression coefficients using PROC GLIMMIX in SAS and used these estimates to make predictions for prevalence in all the counties in the US using RStudio (RStudio Team, 2015).

3.2.2 County Level Prediction

As stated above, we use the estimated random effects to predict random effects for unknown counties via nearest neighboring approach. The random effect for the unknown county is replaced with the random effect of closest county based on Euclidian distance of the centroid of each

$$\tilde{\mu}_{c_i} = \hat{\mu}_{c_j}, s.t. \min \text{dist}(c_i, c_k), k = 1, \dots, m-1,$$

where $\hat{\mu}_{c_j}$is the estimated random effect of county c_j, and c_j is the closest to county c_i. The estimated county level random effects and the model parameters were used to predict the individual level prevalence for E-cigarette current usage and E-cigarette ever usage using the equation (3.2.1).

$$\tilde{\mathrm{P}}_{ijkc}(y_{ijkc} = 1|u) = \frac{\exp(\alpha_i + \beta_j + \gamma_k + x_c'\eta + \mu_c)}{1 + \exp(\alpha_i + \beta_j + \gamma_k + x_c'\eta + \mu_c)}.$$

The county level prevalence was determined using

$$\widetilde{\mathrm{P}}(y_c = 1|u) = \frac{\sum_i \sum_j \sum_k \tilde{P}_{ijkc} \times \mathrm{Pop}_c}{\mathrm{Pop}_{ijkc}},$$

where $\mathrm{Pop}_c = \sum_i \sum_j \sum_k \mathrm{Pop}_{ijkc}$ is total population for county c.

3.4 Results

The model estimates for ever use and current use of E-Cigarettes using FH model for SAE are given in Table (2).

Based on Figure 1, most counties have the prevalence between 21-30 % for every use of E-Cigarettes, followed by prevalence of 11-20 % group.

The predicted and true prevalence for ever-use and current use of E-cigarettes for 142 US counties using FH model is given in Appendix Table 1, and Appendix Table 2, respectively.

Table 3.1. Summary Statistics of prevalence in 143 counties by the total, age group, sex, and race of NYTS 2014-2015 respondents

Characteristic		No. of Respondents E-cigarette Ever	Prevalence E-cig Ever (%)	No. of Respondents E-cig current	Prevalence E-cig current (%)
Total		38683	22.74	38835	10.08
	Missing	746		594	
Age					
	10-14	19846	13.10	19897	5.42
	15-19	18837	32.90	18938	14.98
Sex					
	Male	19553	25.01	19654	11.73
	Female	18969	20.39	19019	8.36
Race					
	White	18189	23.61	16208	11.06
	Black	5781	17.14	5832	5.69
	Hispanic	10598	26.36	10651	11.60
	Others	2357	17.14	2375	8.00

Table 3.2. Model estimates using FH model for ever use and current use of E-Cigarettes

Solutions for Fixed Effects								
	E-Cigarette Ever				E-Cigarette Current			
Effect	Estimate	SE	DF	Pr > |t|	Estimate	SE	DF	Pr > |t|
Intercept	-1.5099	0.16450	129.8	<.0001	-2.4256	0.26570	119.7	<.0001
Age								
10-14	-1.2511	0.00083	36791	<.0001	-1.2317	0.00120	36947	<.0001
15-19	0	.	.	.	0	.	.	.
Gender								
Male	0.2117	0.00071	36791	<.0001	0.3401	0.00097	36947	<.0001
Female	0	.	.	.	0	.	.	.
Race								
White	0.3127	0.00192	36791	<.0001	0.2326	0.00260	36947	<.0001
Black	-0.0618	0.00217	36791	<.0001	-0.4003	0.00307	36947	<.0001
Hispanic	0.5304	0.00199	36791	<.0001	0.3832	0.00271	36947	<.0001
Others	0	.	.	.	0	.	.	.
Poverty (%)	1.8841	1.01710	129.5	0.0662	0.4108	1.64310	119.6	0.8030

Figure 3.1. County prevalence of ever-use of E-Cigarettes in the United States based on 2014-2015 NYTS data

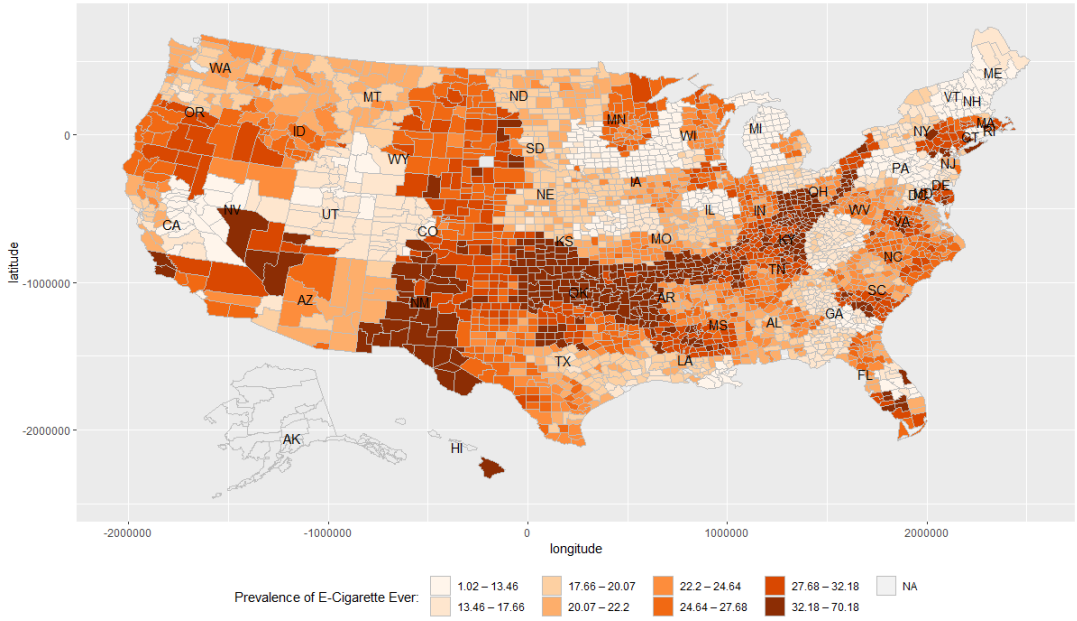
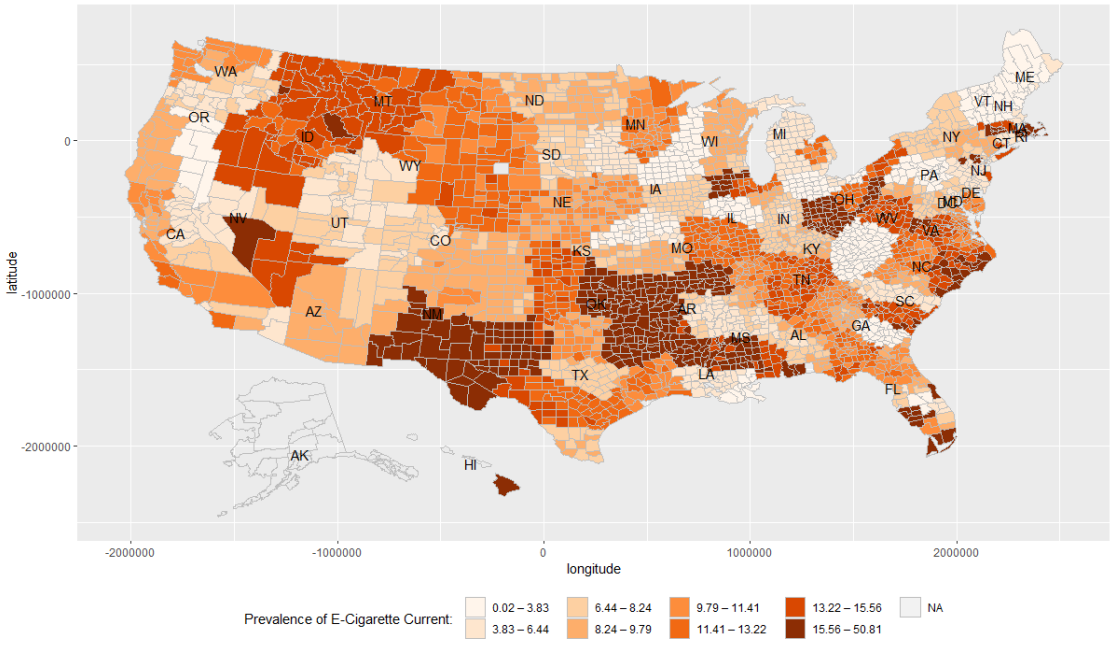


Figure 3.2. County prevalence of current use of E-Cigarettes in the United States based on 2014-2015 NYTS data



Chapter 4. Small Area Estimation using Calibrated Hierarchical Likelihood Approach and Post-Stratitication

4.1 A mixed logistic model for small area estimation

In this chapter, we propose a novel method to estimate binary outcomes in small areas. In 1996, Lee and Nelder extended GLMs to HGLMs by allowing the distribution of random component to be arbitrary conjugates from the exponential family distributions (Lee & Nelder, 1996). For binary outcomes in small area estimation, consider

$$\text{logit } P(y_{ij} = 1) = \boldsymbol{x}_{ij}^t \boldsymbol{\beta} + u_i, \tag{4.1}$$

where $\boldsymbol{y} = (y_{ij}, 1 \le i \le m, 1 \le j \le n_i)$ is the binary response vector which are independent given the random effects $u_1, \dots, u_m$. Let $\boldsymbol{u} = (u_i)_{1 \le i \le m}$, $\boldsymbol{x}_{ij} = \left(x_{ijk}\right)_{1 \le k \le p}$ is a vector of covariates, and $\boldsymbol{\beta} = (\beta_k)_{1 \le k \le p}$ is a vector of unknown fixed effects. The density of $\boldsymbol{y}|\boldsymbol{u}$ follows a Bernoulli distribution with the joint density of $(\boldsymbol{y}, \boldsymbol{u})$ as $f(\boldsymbol{y}, \boldsymbol{u}) = f(\boldsymbol{y}|\boldsymbol{u}) \times f(\boldsymbol{u})$,

$$f(\boldsymbol{y}, \boldsymbol{u}) = \prod_{ij} P(y_{ij} = 1|u_i)^{y_{ij}} \left(1 - P(y_{ij} = 1|u_i)\right)^{1 - y_{ij}} \times f(u_i), \tag{4.2}$$

where $P(y_{ij} = 1) = \exp(\boldsymbol{x}_{ij}^t \boldsymbol{\beta} + u_i) / (1 + \exp(\boldsymbol{x}_{ij}^t \boldsymbol{\beta} + u_i))$, from (4.1). Furthermore, suppose $u_1, \dots, u_m$ are independent and identically distributed as $N(0, \sigma^2)$ with $f(u_i) = (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{1}{2\sigma^2} u_i^2\right)$. This assumption is reasonable to consider as n becomes large, the difference in MHLEs and true values of fixed and random effects converge to a normal distribution, i.e. $(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}, \widehat{\boldsymbol{u}} - \boldsymbol{u}) \to N(0, \tau^2)$ where τ is the limit of the inverse of the observed fisher information matrix ($\boldsymbol{\mathcal{J}}^{-1}$).

Our main goal is to adopt the hierarchical likelihood (h-likelihood also known as h-loglihood) approach to obtain SAE estimation using binary logistic regression. From (4.2), the joint log-density of $\boldsymbol{y}$ and $\boldsymbol{u}$ is

$$\ell = \sum_{ij} \ell_{1ij} + \sum_{i} \ell_{2i}$$

From (4.2), take log,

$$\ell = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left[y_{ij}\log P(y_{ij}=1|u_i) + (1-y_{ij})\log\left(1-P(y_{ij}=1|u_i)\right)\right] + \sum_{i=1}^{m}\log f(u_i)$$

$$= \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left[y_{ij}\log\left(\frac{P(y_{ij}=1)}{1-P(y_{ij}=1)}\right) + \log\left(1-P(y_{ij}=1)\right)\right]$$

$$+ \sum_{i=1}^{m}\log\left[(2\pi\sigma^2)^{-\frac{1}{2}}\exp\left(-\frac{u_i^2}{2\sigma^2}\right)\right]$$

$$= \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left[y_{ij}\log\left(\frac{\left(1+\exp(-\boldsymbol{x}_{ij}^t\boldsymbol{\beta}-u_i)\right)^{-1}}{1-\left(1+\exp(-\boldsymbol{x}_{ij}^t\boldsymbol{\beta}-u_i)\right)^{-1}}\right) + \log\left(1-\left(1+\exp(-\boldsymbol{x}_{ij}^t\boldsymbol{\beta}-u_i)\right)^{-1}\right)\right]$$

$$-\frac{1}{2}\sum_{i=1}^{m}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{m}\log(\sigma^2) + \sum_{i=1}^{m}\log\left[\exp\left(-\frac{u_i^2}{2\sigma^2}\right)\right]$$

$$= \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left[y_{ij}(\boldsymbol{x}_{ij}^t\boldsymbol{\beta}+u_i) - \log(1+\exp(\boldsymbol{x}_{ij}^t\boldsymbol{\beta}+u_i))\right] - \frac{m}{2}\log 2\pi - \frac{m}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{m}u_i^2.$$

Equivalently, we have

$$\ell = \boldsymbol{y}^t(\boldsymbol{X\beta}+\boldsymbol{Zu}) - \boldsymbol{1}^t\log(1+\exp(\boldsymbol{X\beta}+\boldsymbol{Zu})) - \frac{1}{2}\log\left(det(\boldsymbol{G})\right) - \frac{1}{2}\boldsymbol{u}^t\boldsymbol{G}^{-1}\boldsymbol{u} + c,$$

where $c = -\frac{m}{2}\log 2\pi$ is a constant, and $\boldsymbol{G} = \sigma^2\boldsymbol{I}$ is variance-covariance matrix of $\boldsymbol{u}$.

4.2 Hierarchical $(\boldsymbol{h})$ - Likelihood

As stated above, a unique aspect of the *h*-likelihood approach is that it avoids multi-dimensional integration of the nuisance variables when obtaining the parameter estimates, the variables $\boldsymbol{u} = (u_1, \dots, u_m)^t$ are treated as parameters and jointly estimated along with $\boldsymbol{\beta}$, and σ^2. Following Lee and Nelder (2006), the *h*-likelihood for HGLM is given by

$$h = \sum_{ij} \ell_{1ij} + \sum_{i} \ell_{2i} \tag{4.3}$$

where $\ell_{1ij} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left[y_{ij}\left(\boldsymbol{x}_{ij}^t\boldsymbol{\beta} + u_i\right) - \log\left(1 + \left(\boldsymbol{x}_{ij}^t\boldsymbol{\beta} + u_i\right)\right)\right]$ is the logarithm of the conditional likelihood of $\boldsymbol{y}|\boldsymbol{u}$ and $\ell_{2i} = -\frac{m}{2}\log 2\pi - \frac{m}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{m} u_i^2$ is the logarithm of the probability density function of random effect $\boldsymbol{u}$.

4.3 MHLE of $(\boldsymbol{\beta}, \boldsymbol{u})$

Let $\boldsymbol{\tau} = (\boldsymbol{\beta}, \boldsymbol{u})$ be the fixed effect parameters. Given the variance parameter $\boldsymbol{\theta} = \left(\theta_1, \dots, \theta_q\right)^T = \sigma^2$, the maximum hierarchical (h)-likelihood estimators (MHLE) of $\hat{\boldsymbol{\tau}} = \left(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{u}}\right)$ are obtained by solving the score function $\frac{\partial h}{\partial \boldsymbol{\tau}} = 0$. Since the solution does not have a closed form, the MHLE for $(\boldsymbol{\beta}, \boldsymbol{u})$ are obtained using Newton-Raphson approximation through an iterative procedure using equation (2.11).

Holding the variance parameter $\boldsymbol{\theta} = \left(\theta_1, \dots, \theta_q\right)^T = \sigma^2 I_{m\times m}$ constant, the partial derivative of h-likelihood in equation (4.3) with respect to $\boldsymbol{\tau} = (\boldsymbol{\beta}, \boldsymbol{u})$ are

$$h = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left[y_{ij}\left(\boldsymbol{x}_{ij}^t\boldsymbol{\beta} + u_i\right) - \log\left(1 + \exp\left(\boldsymbol{x}_{ij}^t\boldsymbol{\beta} + u_i\right)\right)\right] - \frac{m}{2}\log 2\pi - \frac{m}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{m} u_i^2$$

$$\frac{\partial h}{\partial \beta_r} = \sum_{ij}\left[y_{ij}x_{ijr} - \left(1 + \exp\left(-\boldsymbol{x}_{ij}^t\boldsymbol{\beta} - u_i\right)\right)^{-1}x_{ijr}\right]$$

$$= \sum_{ij}\left[y_{ij}x_{ijr} - p_{ij}x_{ijr}\right],$$

where $p_{ij} = \left(1 + \exp\left(-\boldsymbol{x}_{ij}^t\boldsymbol{\beta} - u_i\right)\right)^{-1}$, $r = 1, \dots, p$ number of fixed covariates. For $i = 1, \dots, m$ random effects, we have

$$\frac{\partial h}{\partial u_i} = \sum_{j=1}^{n_i}\left[y_{ij} - \left(1 + \exp\left(-\boldsymbol{x}_{ij}^t\boldsymbol{\beta} - u_i\right)\right)^{-1}\right] - (\sigma^2)^{-1}u_i$$

$$= \sum_{j=1}^{n_i} [y_{ij} - p_{ij}] - (\sigma^2)^{-1} u_i.$$

The entries of the observed information matrix $\boldsymbol{\mathcal{J}}$ of $\boldsymbol{\beta}$ and $\boldsymbol{u}$ are obtained as follows.

For $r = 1, \dots, p, s = r$,

$$-\frac{\partial^2 h}{\partial \beta_r \partial \beta_s} = \sum_{ij} \left[\exp\left(-\boldsymbol{x}_{ij}^t \boldsymbol{\beta} - u_i\right) \left(1 + \exp\left(-\boldsymbol{x}_{ij}^t \boldsymbol{\beta} - u_i\right)\right)^{-2} x_{ijr} x_{ijs} \right]$$

$$= \sum_{ij} \left[p_{ij}(1 - p_{ij}) x_{ijr} x_{ijs} \right]$$

$$= \sum_{ij} \left[W_{ij} x_{ijr} x_{ijs} \right];$$

for $r = 1, \dots, p, i = 1, \dots, m$,

$$-\frac{\partial^2 h}{\partial \beta_r \partial u_i} = \sum_{j=1}^{n_i} \left[x_{ijr} \exp\left(-\boldsymbol{x}_{ij}^t \boldsymbol{\beta} - u_i\right) \left(1 + \exp\left(-\boldsymbol{x}_{ij}^t \boldsymbol{\beta} - u_i\right)\right)^{-2} \right]$$

$$= \textstyle\sum_{j=1}^{n_i} \left[x_{ijr} p_{ij}(1 - p_{ij}) \right]$$

$$= \sum_{j=1}^{n_i} \left[x_{ijr} W_{ij} \right];$$

for $i = 1, \dots, m, s = 1, \dots, p$,

$$-\frac{\partial^2 h}{\partial u_i \partial \beta_s} = \sum_{j=1}^{n_i} \left[\left(1 + \exp\left(-\boldsymbol{x}_{ij}^t \boldsymbol{\beta} - u_i\right)\right)^{-2} \exp\left(-\boldsymbol{x}_{ij}^t \boldsymbol{\beta} - u_i\right) x_{ijs} \right]$$

$$= \sum_{j=1}^{n_i} \left[p_{ij}(1 - p_{ij}) x_{ijs} \right]$$

$$= \sum_{j=1}^{n_i} \left[W_{ij} x_{ijs} \right];$$

for $i = 1, \dots, m, l = i$,

$$-\frac{\partial^2 h}{\partial u_i \partial u_l} = \sum_{j=1}^{n_i} \left[\left(1 + \exp\left(-\boldsymbol{x}_{ij}^t \boldsymbol{\beta} - u_i\right)\right)^{-2} \exp\left(-\boldsymbol{x}_{ij}^t \boldsymbol{\beta} - u_i\right) \right] + (\sigma^2)^{-1}$$

$$= \sum_{j=1}^{n_i} p_{ij}(1 - p_{ij}) - (\sigma^2)^{-1}$$

$$= \sum_{j=1}^{n_i} W_{ij} - (\sigma^2)^{-1};$$

where $W_{ij} = p_{ij}(1 - p_{ij})$, for $i = 1, \dots, m, l \neq i$, we have $\frac{\partial^2 h}{\partial u_i \partial u_l} = 0$.

Calculating the estimators will be much easier if matrices are used instead of summations. Let $\boldsymbol{y}$ be a $(N \times 1)$ vector of outcome measures, $\boldsymbol{X}$ be a $(N \times p)$ matrix of fixed effects information with p explanatory variables, $\boldsymbol{\beta}$ be a $(p \times 1)$ vector of fixed effects' regression coefficients, $\boldsymbol{Z}$ be the $(N \times m)$ design matrix of explanatory variables with random effects, $\boldsymbol{u}$ be a $(m \times 1)$ vector of random effects, and $\boldsymbol{G} = \sigma^2 I$ be the $(m \times m)$ variance-covariance matrix of random effects.

The h-likelihood estimation of $\boldsymbol{\beta}$ and $\boldsymbol{u}$ are obtained via Newton Raphson iterative algorithm using the equation (2.11)

$$\begin{pmatrix} \widehat{\boldsymbol{\beta}}^{(k+1)} \\ \widehat{\boldsymbol{u}}^{(k+1)} \end{pmatrix} = \begin{pmatrix} \widehat{\boldsymbol{\beta}}^{(k)} \\ \widehat{\boldsymbol{u}}^{(k)} \end{pmatrix} + \left(\boldsymbol{\mathcal{J}}^{-1}\boldsymbol{S}(\tau)\right)\Big|_{(\boldsymbol{\beta},\boldsymbol{u})=(\widehat{\boldsymbol{\beta}}^{(k)},\widehat{\boldsymbol{u}}^{(k)})} \tag{4.4}$$

$$= \begin{pmatrix} \widehat{\boldsymbol{\beta}}^{(k)} \\ \widehat{\boldsymbol{u}}^{(k)} \end{pmatrix} + \begin{pmatrix} -\frac{\partial^2 h}{\partial \boldsymbol{\beta}^2} & -\frac{\partial^2 h}{\partial \boldsymbol{\beta} \partial \boldsymbol{u}} \\ -\frac{\partial^2 h}{\partial \boldsymbol{u} \partial \boldsymbol{\beta}} & -\frac{\partial^2 h}{\partial \boldsymbol{u}^2} \end{pmatrix}^{-1} \begin{pmatrix} \frac{\partial h}{\partial \boldsymbol{\beta}} \\ \frac{\partial h}{\partial \boldsymbol{u}} \end{pmatrix}\Bigg|_{(\boldsymbol{\beta},\boldsymbol{u})=(\widehat{\boldsymbol{\beta}}^{(k)},\widehat{\boldsymbol{u}}^{(k)})},$$

by replacing $\boldsymbol{\beta}$, $\boldsymbol{u}$ in $\boldsymbol{\mathcal{J}}$ and $\boldsymbol{S}(\tau)$ with $\widehat{\boldsymbol{\beta}}^{(k)}$ and $\widehat{\boldsymbol{u}}^{(k)}$ respectively. The variance-covariance matrix $\boldsymbol{\theta}$ will be estimated by MHLE of $\boldsymbol{\theta}$ iteratively as described in section 4.5 via adjusted h-likelihood (h_A) approach.

4.4 Bias Correction

As stated above, random effects and fixed effects are jointly estimated using HMLE procedure through Newton Raphson approximation. However, it is shown that directly plugging estimated $\boldsymbol{u}$, $\widehat{\boldsymbol{u}}$ which was estimated using h-likelihood will result in biasness when estimating the parameters of interest, specially based on nonlinear likelihood functions. Moreover, MHLE of $\widehat{\boldsymbol{u}}$

will lead to bias and inconsistent estimations of regression coefficients and parameters in the random effects, such as variance parameters. The researchers have considered and proposed many methods to avoid this biasness in the parameter estimations (Carroll, Ruppert, Stefanski, & Crainiceanu, 2006; Hausman, Newey, & Powell, 1995; C. Wang, Hsu, Feng, & Prentice, 1997). We use regression calibration approach proposed by Wang et.al (1997) to correct for biasness in the estimators. The use of correction factor is a simpler approach to correct the biasness of MLE, which MLE is adjusted based on the sample size. Furthermore, the regression calibration method is computationally efficient and has been applied widely in correcting measurement errors in many applications since it provides reliable estimates than direct estimates.

Wang et. al (1997) used the idea from the large sample theory and proposed the regression calibration method for measurement error correction. The main idea of this method is to adjust $\widehat{\boldsymbol{u}}$ using variance of $\boldsymbol{u}|\widehat{\boldsymbol{u}}$. From the large sample theory, when *N* becomes large, maximum likelihood estimators of likelihood function has the asymptotic normal property $\sqrt{N}(\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}) \approx N\left(0, \left(\boldsymbol{\mathcal{J}}(\boldsymbol{\tau})\right)^{-1}\right)$, where $\boldsymbol{\tau} = (\boldsymbol{\beta}, \boldsymbol{u})$, $\hat{\boldsymbol{\tau}}$ is consistent and asymptotically efficient estimator of $\boldsymbol{\tau}$, which is also known as the Cramer-Rao Lower Bound(Casella & Berger, 2002)**.** The variance of $\boldsymbol{u}|\widehat{\boldsymbol{u}}$, $\boldsymbol{\gamma}$ is the $(m \times m)$ right lower corner matrix of $\left(\boldsymbol{\mathcal{J}}(\boldsymbol{\tau})\right)^{-1}$is the asymptotic variance covariance matrix of maximum likelihood estimates.

Since $\boldsymbol{u} \sim N(0, \sigma^2)$, $(\boldsymbol{u}, \widehat{\boldsymbol{u}})$ has a joint normal distribution,

$$\begin{pmatrix} \boldsymbol{u} \\ \widehat{\boldsymbol{u}} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \sigma^2 \\ \sigma^2 & \sigma^2 + \boldsymbol{\gamma}^2 \end{pmatrix}\right).$$

Then, the conditional distribution of $\boldsymbol{u}$ given $\widehat{\boldsymbol{u}}$ can be derived as

$$\boldsymbol{u}|\widehat{\boldsymbol{u}} \sim N\left(\boldsymbol{\zeta}\widehat{\boldsymbol{u}}, \sigma^2(1 - \boldsymbol{\zeta})\right), \text{ where } \boldsymbol{\zeta} = \frac{\sigma^2}{\sigma^2 + \boldsymbol{\gamma}^2}, \boldsymbol{\gamma} \text{ is the limit of } \boldsymbol{\mathcal{J}}^{-1}.$$

Now, under regression calibration method, we replace $\boldsymbol{u}$ by

$$E[\boldsymbol{u}|\widehat{\boldsymbol{u}}] = \boldsymbol{\zeta}\widehat{\boldsymbol{u}} = \frac{\sigma^2}{\sigma^2 + \boldsymbol{\gamma}^2}\widehat{\boldsymbol{u}}.$$

Note that, if the variance of $\widehat{\boldsymbol{u}}$, $\boldsymbol{\gamma} \approx 0$, then $E[\boldsymbol{u}|\widehat{\boldsymbol{u}}] \approx 1$, that means the regression calibration methods and original MHLE estimates are similar. However, when $\mathrm{Var}(\widehat{\boldsymbol{u}})$ is large, the MHLE result the biased estimates, hence regression calibration method adjusts the biasness of the estimates.

4.5 MHLE of variance parameter $(\boldsymbol{\theta})$

4.5.1 Iterative approach based on the partial derivative

From equation (2.13), the adjusted h-likelihood

$$h_A = h|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}},\boldsymbol{u}=\widehat{\boldsymbol{u}}} - \frac{1}{2}\log\{\det(\boldsymbol{\mathcal{J}})\}|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}},\boldsymbol{u}=\widehat{\boldsymbol{u}}} + \frac{p+m}{2}\log 2\pi,$$

where m is dispersion parameters, and p is the number of fixed effects.

The generic way of obtaining the maximum adjusted profile h-likelihood estimator for θ is by solving the score function $\partial h_A/\partial\theta = 0$, but this becomes complicated when there is no closed form for the θ. Thus, it requires some numerical approximation. First, consider the score function of the adjusted profile h-likelihood from the equation (2.14)

$$\frac{\partial h_A}{\partial\theta} = \frac{\partial h|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}},\boldsymbol{u}=\widehat{\boldsymbol{u}}}}{\partial\theta} - \frac{1}{2}\mathrm{trace}\left(\boldsymbol{\mathcal{J}}^{-1}\frac{\partial\boldsymbol{\mathcal{J}}}{\partial\theta}\right)\Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}},\boldsymbol{u}=\widehat{\boldsymbol{u}}}, \tag{4.5}$$

is used to estimate the variance-covariance matrix $\boldsymbol{\theta}$ of random effect $\boldsymbol{u}$. Since $\partial\ell_{1ij}/\partial\theta = 0$, the score function can be written as

$$\frac{\partial h}{\partial\theta} = 0 + \frac{\partial}{\partial\theta}\left(\sum_{i=1}^{m}\ell_{2i}\right)$$

The adjusted profile log likelihood of $u_i \sim N(0,\theta)$ for small area i

$$\ell_{2i} = -\frac{m}{2}\log 2\pi - \frac{m}{2}\log\theta - \frac{1}{2\theta}\sum_{i=1}^{m}u_i^2 = -\frac{m}{2}\log 2\pi - \frac{m}{2}\log\theta - \frac{1}{2\theta}\boldsymbol{u}^T\boldsymbol{u},$$

where $\theta I_{m\times m} = \boldsymbol{G} = \sigma^2 I_{m\times m}$. Now, from $h = \sum_{ij}\ell_{1ij} + \sum_i \ell_{2i}$, $\partial h/\partial\theta$ given $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}, \boldsymbol{u} = \widehat{\boldsymbol{u}}$ can be represented as

$$\frac{\partial h|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}},\boldsymbol{u}=\widehat{\boldsymbol{u}}}}{\partial\theta} = 0 + \frac{\partial}{\partial\theta}\left(-\frac{m}{2}\log\theta - \frac{1}{2\theta}\sum_{i=1}^{m}\hat{u}_i^2\right)$$

$$= \left(-\frac{m}{2\theta} + \frac{1}{2\theta^2} \sum_{i=1}^{m} \hat{u}_i^2 \right) \Bigg|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}, \boldsymbol{u}=\widehat{\boldsymbol{u}}} .$$

The partial derivative of the observed information matrix with respect to θ given $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}, \boldsymbol{u} = \widehat{\boldsymbol{u}}$

$$\frac{\partial \boldsymbol{\mathcal{J}}}{\partial \theta}\Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}, \boldsymbol{u}=\widehat{\boldsymbol{u}}} = \frac{\partial}{\partial \theta} \begin{pmatrix} \left(-\frac{\partial^2 h}{\partial \boldsymbol{\beta}^2} \right)_{p \times p} & \left(-\frac{\partial^2 h}{\partial \boldsymbol{\beta} \partial \boldsymbol{u}} \right)_{p \times m} \\ \left(-\frac{\partial^2 h}{\partial \boldsymbol{u} \partial \boldsymbol{\beta}} \right)_{m \times p} & \left(-\frac{\partial^2 h}{\partial \boldsymbol{u}^2} \right)_{m \times m} \end{pmatrix},$$

From (4.5)

$$\frac{\partial h_A}{\partial \theta} = \left(-\frac{m}{2\theta} + \frac{1}{2\theta^2} \sum_{i=1}^{m} \hat{u}_i^2 \right) \Bigg|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}, \boldsymbol{u}=\widehat{\boldsymbol{u}}}$$

$$- \frac{1}{2} \operatorname{trace} \left(\begin{pmatrix} \boldsymbol{\mathcal{J}}_{11} & \boldsymbol{\mathcal{J}}_{12} \\ \boldsymbol{\mathcal{J}}_{21} & \boldsymbol{\mathcal{J}}_{22} \end{pmatrix}^{-1} \frac{\partial}{\partial \theta} \begin{pmatrix} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X} & \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{Z} \\ \boldsymbol{Z}^T \boldsymbol{W} \boldsymbol{X} & \boldsymbol{Z}^T \boldsymbol{W} \boldsymbol{Z} + (\theta I_{m \times m})^{-1} \end{pmatrix} \right) \Bigg|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}, \boldsymbol{u}=\widehat{\boldsymbol{u}}}$$

$$= \left(-\frac{m}{2\theta} + \frac{1}{2\theta^2} \sum_{i=1}^{m} \hat{u}_i^2 \right) \Bigg|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}, \boldsymbol{u}=\widehat{\boldsymbol{u}}} - \frac{1}{2} \operatorname{trace} \left(\begin{pmatrix} \boldsymbol{\mathcal{J}}_{11} & \boldsymbol{\mathcal{J}}_{12} \\ \boldsymbol{\mathcal{J}}_{21} & \boldsymbol{\mathcal{J}}_{22} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & -\theta^{-2} I_{m \times m} \end{pmatrix} \right) \Bigg|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}, \boldsymbol{u}=\widehat{\boldsymbol{u}}}$$

$$= \left(-\frac{m}{2\theta} + \frac{1}{2\theta^2} \sum_{i=1}^{m} \hat{u}_i^2 \right) \Bigg|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}, \boldsymbol{u}=\widehat{\boldsymbol{u}}} - \frac{1}{2} \operatorname{trace} \left(\begin{pmatrix} \boldsymbol{\mathcal{J}}_{11}^* & \boldsymbol{\mathcal{J}}_{12}^* \\ \boldsymbol{\mathcal{J}}_{21}^* & \boldsymbol{\mathcal{J}}_{22}^* \end{pmatrix} \begin{pmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & -\theta^{-2} I_{m \times m} \end{pmatrix} \right) \Bigg|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}, \boldsymbol{u}=\widehat{\boldsymbol{u}}}$$

$$= \left(-\frac{m}{2\theta} + \frac{1}{2\theta^2} \sum_{i=1}^{m} \hat{u}_i^2 \right) \Bigg|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}, \boldsymbol{u}=\widehat{\boldsymbol{u}}} - \frac{1}{2} \operatorname{trace} \begin{pmatrix} \boldsymbol{0} & -\boldsymbol{\mathcal{J}}_{12}^* \theta^{-2} I_{m \times m} \\ \boldsymbol{0} & -\boldsymbol{\mathcal{J}}_{22}^* \theta^{-2} I_{m \times m} \end{pmatrix} \Bigg|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}, \boldsymbol{u}=\widehat{\boldsymbol{u}}}$$

$$= \left(-\frac{m}{2\theta} + \frac{1}{2\theta^2} \sum_{i=1}^{m} \hat{u}_i^2 \right) \Bigg|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}, \boldsymbol{u}=\widehat{\boldsymbol{u}}} + \frac{1}{2\theta^2} \operatorname{trace}(\boldsymbol{\mathcal{J}}_{22}^*) \Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}, \boldsymbol{u}=\widehat{\boldsymbol{u}}}$$

Set $\partial h_A / \partial \theta = 0$

$$\left(-\frac{m}{2\theta} + \frac{1}{2\theta^2} \sum_{i=1}^{m} \hat{u}_i^2 \right) \Bigg|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}, \boldsymbol{u}=\widehat{\boldsymbol{u}}} + \frac{1}{2\theta^2} \operatorname{trace}(\boldsymbol{\mathcal{J}}_{22}^*) \Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}, \boldsymbol{u}=\widehat{\boldsymbol{u}}} = 0.$$

Thus,

$$\hat{\theta} = \frac{1}{m}\left(\sum_{i=1}^{m} \hat{u}_i^2\right)\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}},\boldsymbol{u}=\hat{\boldsymbol{u}}} + \frac{1}{m}\text{trace}(\boldsymbol{\mathcal{J}}_{22}^{*})|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}},\boldsymbol{u}=\hat{\boldsymbol{u}}}, \tag{4.6}$$

where $\hat{\theta}$ is the MLE of the variance parameter θ. p is the number of fixed effects, $\boldsymbol{\mathcal{J}}_{22}^{*}$ is the lower right block matrix of $\boldsymbol{\mathcal{J}}^{-1}$ with dimensions being $m \times m$ which is calculated using current estimates of $\boldsymbol{\beta}$ and $\boldsymbol{u}$. As described in appendix C, by the asymptotic properties of MHLE of $\boldsymbol{u}$, $\boldsymbol{\mathcal{J}}_{22}^{*} = (-\partial^2 h|\partial \boldsymbol{u}^2|_{\boldsymbol{u}=\hat{\boldsymbol{u}}})^{-1}$.

Algorithm

1. Set $\boldsymbol{\beta}_0, \boldsymbol{u}_0$, and $\theta_0(\sigma_0^2)$.
2. Evaluate quantities $\boldsymbol{S}(\boldsymbol{\tau})$ and $\boldsymbol{\mathcal{J}}$.
3. Estimate $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}^{(t)}$, and $\boldsymbol{u} = \widehat{\boldsymbol{u}}^{(t)}$using Newton Raphson $\big(\text{eqn. (4)}\big)$.
4. Update $\boldsymbol{\mathcal{J}}$ using $\widehat{\boldsymbol{\beta}}^{(t)}$and $\widehat{\boldsymbol{u}}^{(t)}$ and estimate $\boldsymbol{\mathcal{J}}_{22}^{*}$ using $\boldsymbol{\mathcal{J}}^{-1}$ matrix.
5. Estimate $\hat{\theta}^{(t)}$ using equation (7), by replacing $\boldsymbol{u}$ with $E\big[\boldsymbol{u}\big|\widehat{\boldsymbol{u}}^{(t)}\big] = \hat{\boldsymbol{\zeta}}^{(t)}\widehat{\boldsymbol{u}}^{(t)} = \widehat{\boldsymbol{u}}^{c(t)}$, where $\boldsymbol{\zeta} = (\zeta_1, \dots, \zeta_m)$, $\hat{\boldsymbol{\zeta}}^{(t)} = \frac{\hat{\theta}^{(t-1)}}{\hat{\theta}^{(t-1)} + \boldsymbol{\tau}_i^{(t)}}$, $\hat{\theta}^{(t-1)} = \text{Var}\big(\widehat{\boldsymbol{u}}^{(t-1)}\big)$, $\boldsymbol{\tau}_i^{(t)} = \boldsymbol{\mathcal{J}}_{ii}^{*(t)}$, $\boldsymbol{\mathcal{J}}_{ii}^{*}$ is the (i,i) diagonal element of the lower right corner matrix of 2×2 $\boldsymbol{\mathcal{J}}^{-1}$ block matrix $((\boldsymbol{\mathcal{J}}^{-1})_{22})$ evaluated at t^{th} iteration. By asymptotic properties of $\widehat{\boldsymbol{u}}$, $(\boldsymbol{\mathcal{J}}^{-1})_{22} = (\boldsymbol{\mathcal{J}}_{22})^{-1}$.
6. Update h by replacing $\boldsymbol{u}$ with $\widehat{\boldsymbol{u}}^{c(t)}$, and $\exp \boldsymbol{u}$ with $E\big[\exp \boldsymbol{u}\,\big|\widehat{\boldsymbol{u}}^{(t)}\big] = \exp\Big(\hat{\boldsymbol{\zeta}}^{(t)}\widehat{\boldsymbol{u}}^{(t)} + \Big(\hat{\theta}^{(t)}\big(1 - \hat{\boldsymbol{\zeta}}^{(t)}\big)\Big)/2\Big)$.
7. Repeat step 2 to 6 until it meets the convergence criteria, which is defined as

$$\max\{|\widehat{\boldsymbol{\beta}}^{(i+1)} - \widehat{\boldsymbol{\beta}}^{(i)}|, |\hat{\theta}^{(i+1)} - \hat{\theta}^{(i)}|\} < \delta$$

where δ is a predetermined tolerance limit. The MHLEs of $\boldsymbol{\beta}, \boldsymbol{u}$, and σ are $\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{u}}$, and $\hat{\sigma}^2\big(= \hat{\theta}^{(t)}\big)$, respectively.

4.6 Asymptotic Properties of MHLE of $\widehat{\boldsymbol{\beta}}$

The Taylor series expansion of $\partial h/\partial \beta_i$ around $u = \hat{u}$,

$$\frac{\partial h}{\partial \beta_i} = \frac{\partial h}{\partial \beta_i}\Big|_{u_i=\hat{u}_i} + (u_i - \hat{u}_i)C_1|_{u_i=\hat{u}_i} + \frac{1}{2!}(u_i - \hat{u}_i)^T C_2 (u_i - \hat{u}_i)\Big|_{u_i=\hat{u}_i} + \cdots$$

where $C_1 = (\partial/\partial\beta_i)(\partial h/\partial u_i)|_{u_i=\hat{u}_i}$, and $C_2 = (\partial/\partial\beta_i)\left(\partial^2 h/\partial u_i^2\right)\Big|_{u_i=\hat{u}_i}$.

From $(A.2)$, $n(E(u_i|y) - \hat{u}_i) = n\left(\mathcal{O}(n^{-1})\right) = \mathcal{O}(1)$, then,

$$E\left[C_1(u_i - \hat{u}_i)|_{u_i=\hat{u}_i}\right] = E\left[\frac{C_1}{n}\mathcal{O}(1)\right] = \frac{C_1}{n} = \mathcal{O}(1).$$

Similarly, from $(A.4)$,

$$E\left[(u_i - \hat{u}_i)^T C_2 (u_i - \hat{u}_i)|_{u_i=\hat{u}_i}\right] = \frac{C_2}{n} = \mathcal{O}(1).$$

Now,

$$E\left[\frac{\partial h}{\partial \beta_i}\right] = \frac{\partial h}{\partial \beta_i}\Big|_{u_i=\hat{u}_i} + E\left[C_1(u_i - \hat{u}_i)|_{u_i=\hat{u}_i}\right] + \frac{1}{2!}E\left[(u_i - \hat{u}_i)^T C_2 (u_i - \hat{u}_i)|_{u_i=\hat{u}_i}\right] + \cdots$$

$$= \frac{\partial h}{\partial \beta_i}\Big|_{u_i=\hat{u}_i} + \mathcal{O}(1),$$

where $C_j/n = \mathcal{O}(1),\ j = 1,2.$

4.7 Asymptotic Properties of MHLE of $\hat{\boldsymbol{u}}$

First, consider the second-order Taylor Series expansion of the joint log-likelihood, similarly h_i around $u_i = \hat{u}_i$, then the numerator of $E[u_i|y]$

$$\int u_i \exp h(u_i)\, du_i = \int u_i \exp\left\{h(\hat{u}_i) + \frac{1}{2}(u_i - \hat{u}_i)^T h''(u_i)|_{u_i=\hat{u}_i}(u_i - \hat{u}_i) + \cdots\right\} du_i$$

$$= \exp\left(h(\hat{u}_i)\right)\left\{\int (u_i - \hat{u}_i)\exp\left\{\frac{1}{2}(u_i - \hat{u}_i)^T h''(u_i)|_{u_i=\hat{u}_i}(u_i - \hat{u}_i)\right\} du_i \right.$$

$$\left. + \int \hat{u}_i \exp\left\{\frac{1}{2}(u_i - \hat{u}_i)^T h''(u_i)|_{u_i=\hat{u}_i}(u_i - \hat{u}_i)\right\} du_i\right\}.$$

Note that,

$$\int (u_i - \hat{u}_i)\exp\left\{\frac{1}{2}(u_i - \hat{u}_i)^T h''(u_i)|_{u_i=\hat{u}_i}(u_i - \hat{u}_i)\right\} du_i = 0 \text{ at } u_i = \hat{u}_i.$$

Thus,

$$\int u_i \exp h(u_i)\, du_i = \hat{u}_i \exp\big(h(\hat{u}_i)\big) \exp\left\{\frac{1}{2}(u_i - \hat{u}_i)^T h''(u_i)|_{u_i=\hat{u}_i}(u_i - \hat{u}_i)\right\} du_i$$

$$= \hat{u}_i \left|-2\pi\big(h''(u_i)|_{u_i=\hat{u}_i}\big)^{-1}\right|^{1/2} \exp\{h(\hat{u}_i)\}.$$

Similarly, the denominator of $E[u_i|y]$

$$\int \exp h(u_i)\, du_i \approx \int \exp\left\{h(\hat{u}_i) + \frac{1}{2}(u_i - \hat{u}_i)^T h''(u_i)|_{u_i=\hat{u}_i}(u_i - \hat{u}_i) + \cdots\right\} du_i,$$

$$= \exp\big(h(\hat{u}_i)\big) \int \exp\left\{\frac{1}{2}(u_i - \hat{u}_i)^T h''(u_i)|_{u_i=\hat{u}_i}(u_i - \hat{u}_i)\right\} du_i\,,$$

$$= \left|-2\pi\big(h''(u_i)|_{u_i=\hat{u}_i}\big)^{-1}\right|^{1/2} \exp\big(h(\hat{u}_i)\big)$$

Now, by taking the ratio

$$E[u_i|y] = \hat{u}_i, \tag{4.7}$$

which is the first order Laplace Approximation to the $u_i|y$. The asymptotic order of n terms of the Taylor Series expansion has a relative error of $\mathcal{O}(n^{-1})$ where $\lim_{n\to\infty} \mathcal{O}(n^{-1}) \approx \epsilon$, hence, the conditional mean of u_i given y,

$$E(u_i|y) = \hat{u}_i + \mathcal{O}(n^{-1}) \tag{4.8}$$

Next, consider the $E\big[(u_i^2|y)\big]$,

$$E\big[(u_i^2|y)\big] = \frac{\int u_i^2 \exp h(u_i)\, du_i}{\int \exp h(u_i)\, du_i}$$

$$\approx \frac{\int u_i^2 \exp\left\{h(\hat{u}_i) + \frac{1}{2}(u_i - \hat{u}_i)^T h''(u_i)|_{u_i=\hat{u}_i}(u_i - \hat{u}_i) + \cdots\right\} du_i}{\left(-2\pi\big(h''(u_i)|_{u_i=\hat{u}_i}\big)^{-1}\right)^{\frac{1}{2}} \exp\big(h(\hat{u}_i)\big)}$$

$$= \left|-\frac{1}{2\pi} h''(u_i)|_{u_i=\hat{u}_i}\right|^{\frac{1}{2}} \int u_i^2 \exp\left\{\frac{1}{2}(u_i - \hat{u}_i)^T h''(u_i)|_{u_i=\hat{u}_i}(u_i - \hat{u}_i)\right\} du_i$$

$$= \left|-\frac{1}{2\pi} h''(u_i)|_{u_i=\hat{u}_i}\right|^{\frac{1}{2}} \left(\int (u_i - \hat{u}_i)^2 \exp\left\{\frac{1}{2}(u_i - \hat{u}_i)^T h''(u_i)|_{u_i=\hat{u}_i}(u_i - \hat{u}_i)\right\} du_i \right.$$

$$- \int \hat{u}_i{}^2 \exp\left\{\frac{1}{2}(u_i - \hat{u}_i)^T h''(u_i)|_{u_i=\hat{u}_i}(u_i - \hat{u}_i)\right\} du_i$$

$$\left. + 2\int u_i \hat{u}_i \exp\left\{\frac{1}{2}(u_i - \hat{u}_i)^T h''(u_i)|_{u_i=\hat{u}_i}(u_i - \hat{u}_i)\right\} du_i \right)$$

$$= \left|-\frac{1}{2\pi} h''(u_i)|_{u_i=\hat{u}_i}\right|^{\frac{1}{2}} \left(\left(h''(u_i)|_{u_i=\hat{u}_i}\right)^{-1} \left|-\frac{1}{2\pi} h''(u_i)|_{u_i=\hat{u}_i}\right|^{-\frac{1}{2}} \right.$$

$$\left. + \hat{u}_i{}^2 \left|-\frac{1}{2\pi} h''(u_i)|_{u_i=\hat{u}_i}\right|^{-\frac{1}{2}} \right)$$

$$E\left[\left(u_i^2 | y\right)\right] = \left(h''(u_i)|_{u_i=\hat{u}_i}\right)^{-1} + \hat{u}_i{}^2, \tag{4.9}$$

where $h''(u_i)|_{u_i=\hat{u}_i} = \left.\frac{\partial^2 h}{\partial u_i}\right|_{u_i=\hat{u}_i}$. Now the conditional variance of $u_i|y$ from (4.8) and (4.9),

$$\mathrm{Var}(u_i|y) = \left(-\partial^2 h | \partial u_i^2 \big|_{u_i=\hat{u}_i}\right)^{-1} + \mathcal{O}(n^{-1}). \tag{4.10}$$

The conditional variance $u_i|y$ can be simplified as $\lim_{n\to\infty} \mathrm{Var}(u_i|y) = \left(-\partial^2 h | \partial u_i^2 \big|_{u_i=\hat{u}_i}\right)^{-1}$.

Chapter 5. Penalized CPH with Bias Correction in small area estimation

Penalized h-likelihood for sufficiently large m

$$h = \ell_1 + \ell_2 - p_\gamma(|\boldsymbol{u}|),$$

where $p_\gamma(|\boldsymbol{u}|)$ is the penalty function for random effects $\boldsymbol{u}$, $p_\gamma^{L_1}(|\boldsymbol{u}|)$ is the LASSO L_1 penalty of random effects,

$$p_\gamma(|\boldsymbol{u}|) = p_\gamma^{L_1}(|\boldsymbol{u}|) = \gamma|\boldsymbol{u}|,$$

γ is the tuning parameter which reduces the complexity of the model when m is large by letting u go to zero, thus reducing the dimensionality of random effects (Friedman, Hastie, & Tibshirani, 2001; She, 2009; Tibshirani, 1996, 2011). We consider the LASSO penalty, $p_\gamma(|\boldsymbol{u}|) = \gamma|\boldsymbol{u}|$. This will remove the null random effects from the model; hence it improves the prediction performance.

$$h = \boldsymbol{y}^T(\boldsymbol{X\beta} + \boldsymbol{Zu}) - \boldsymbol{1}^T \log(1 + \exp(\boldsymbol{X\beta} + \boldsymbol{Zu})) - \frac{m}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\boldsymbol{u}^T\boldsymbol{u} - \boldsymbol{1}^T\left[p_\gamma(|\boldsymbol{u}|)\right] + c,$$

$$\frac{\partial h}{\partial \boldsymbol{\beta}} = \left[\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{X}^T(1 + \exp(\boldsymbol{X\beta} + \boldsymbol{Zu}))^{-1}\exp(\boldsymbol{X\beta} + \boldsymbol{Zu})\right]_{p\times 1}$$

$$= \boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{X}^T\boldsymbol{\pi} = \boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{\pi}),$$

$$\frac{\partial h}{\partial \boldsymbol{u}} = \left[\boldsymbol{Z}^T\boldsymbol{y} - \boldsymbol{Z}^T(1 + \exp(\boldsymbol{X\beta} + \boldsymbol{Zu})^{-1}\exp(\boldsymbol{X\beta} + \boldsymbol{Zu}) - \boldsymbol{G}^{-1}\boldsymbol{u} - \left[p_\gamma(|\boldsymbol{u}|)\right]'\right]_{m\times 1}$$

$$= \boldsymbol{Z}^T\boldsymbol{y} - \boldsymbol{Z}^T\boldsymbol{\pi} - \boldsymbol{G}^{-1}\boldsymbol{u} - \left[p_\gamma(|\boldsymbol{u}|)\right]' = \boldsymbol{Z}^T(\boldsymbol{y} - \boldsymbol{\pi}) - \boldsymbol{G}^{-1}\boldsymbol{u} - \left[p_\gamma(|\boldsymbol{u}|)\right]',$$

where $\boldsymbol{\pi} = 1/(1 + \exp(-\boldsymbol{X\beta} - \boldsymbol{Zu}))$.

The components of the asymptotic variance-covariance matrix of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{u}}$, $\boldsymbol{\mathcal{J}}$ also known as the observed information matrix that is calculated by

$$\boldsymbol{\mathcal{J}} = \begin{bmatrix} -\dfrac{\partial^2 h}{\partial \boldsymbol{\beta}^2} & -\dfrac{\partial^2 h}{\partial \boldsymbol{\beta}\partial \boldsymbol{u}} \\ -\dfrac{\partial^2 h}{\partial \boldsymbol{u}\partial \boldsymbol{\beta}} & -\dfrac{\partial^2 h}{\partial \boldsymbol{u}^2} \end{bmatrix},$$

where

$$-\frac{\partial^2 h}{\partial \boldsymbol{\beta}^2} = \frac{\partial}{\partial \boldsymbol{\beta}}(\boldsymbol{X}^T(1+\exp(-\boldsymbol{X\beta}-\boldsymbol{Zu}))^{-1}\exp(-\boldsymbol{X\beta}-\boldsymbol{Zu})\boldsymbol{X})$$

$$= \boldsymbol{X}^T \frac{1}{(1+\exp(-\boldsymbol{X\beta}-\boldsymbol{Zu}))}\left(\frac{1}{1+\exp(\boldsymbol{X\beta}+\boldsymbol{Zu})}\right)\boldsymbol{X}$$

$$= \boldsymbol{X}^T\boldsymbol{\pi}(\mathbf{1}-\boldsymbol{\pi})\boldsymbol{X}$$

$$= \boldsymbol{X}^T\boldsymbol{WX},$$

$$-\frac{\partial^2 h}{\partial \boldsymbol{\beta}\partial \boldsymbol{u}} = \boldsymbol{X}^T\frac{\partial}{\partial \boldsymbol{u}}((1+\exp(-\boldsymbol{X\beta}-\boldsymbol{Zu}))^{-1}\exp(-\boldsymbol{X\beta}-\boldsymbol{Zu}))$$

$$= \boldsymbol{X}^T(1+\exp(-\boldsymbol{X\beta}-\boldsymbol{Zu}))^{-2}\boldsymbol{Z}$$

$$= \boldsymbol{X}^T\boldsymbol{\pi}(\mathbf{1}-\boldsymbol{\pi})\boldsymbol{Z}$$

$$= \boldsymbol{X}^T\boldsymbol{WZ},$$

$$-\frac{\partial^2 h}{\partial \boldsymbol{u}\partial \boldsymbol{\beta}} = -\frac{\partial}{\partial \boldsymbol{\beta}}(\boldsymbol{Z}^T\boldsymbol{y}-\boldsymbol{Z}^T(1+\exp(-\boldsymbol{X\beta}-\boldsymbol{Zu}))^{-1}-\boldsymbol{G}^{-1}\boldsymbol{u})$$

$$= \boldsymbol{Z}^T(1+\exp(-\boldsymbol{X\beta}-\boldsymbol{Zu}))^{-2}\boldsymbol{X}$$

$$= \boldsymbol{Z}^T\boldsymbol{\pi}(\mathbf{1}-\boldsymbol{\pi})\boldsymbol{X}$$

$$= \boldsymbol{Z}^T\boldsymbol{WX},$$

$$\frac{\partial h}{\partial \boldsymbol{u}} = \left[\boldsymbol{Z}^T\boldsymbol{y}-\boldsymbol{Z}^T(1+\exp(\boldsymbol{X\beta}+\boldsymbol{Zu})^{-1}\exp(\boldsymbol{X\beta}+\boldsymbol{Zu})-\boldsymbol{G}^{-1}\boldsymbol{u}-\left[p_\gamma(|\boldsymbol{u}|)\right]'\right]_{m\times 1}$$

$$\left[p_\gamma(|\boldsymbol{u}|)\right]' = \frac{\partial(\gamma|\boldsymbol{u}|)}{\partial \boldsymbol{u}} = \gamma\,(|\boldsymbol{u}|)\boldsymbol{u}^{-1}$$

The penalty function $p_\gamma(|\boldsymbol{u}|)$ is non-differentiable at the origin and the second derivative $[p_\gamma(|\boldsymbol{u}|)]''$ does not exist, hence $[p_\gamma(|\boldsymbol{u}|)]$ is approximated based on a local quadratic approximation using

$$[p_\gamma(|\boldsymbol{u}|)]' = p'_\gamma(|\boldsymbol{u}|)\text{sgn}(|\boldsymbol{u}|) \approx \left\{\frac{p'_\gamma(|\boldsymbol{u}^{(0)}|)}{\boldsymbol{u}^{(0)}}\right\}\boldsymbol{u}, \qquad \text{for } \boldsymbol{u} = \boldsymbol{u}^{(0)}.$$

Now, take the second order partial derivative of h-likelihood with respect to $\boldsymbol{u}$,

$$-\frac{\partial^2 h}{\partial \boldsymbol{u}^2} = -\frac{\partial}{\partial \boldsymbol{u}}\left(\boldsymbol{Z}^T\boldsymbol{y} - \boldsymbol{Z}^T(1+\exp(-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{Z}\boldsymbol{u}))^{-1} - \boldsymbol{G}^{-1}\boldsymbol{u} - p'_\gamma(|\boldsymbol{u}|)\text{sgn}(|\boldsymbol{u}|)\right)$$

$$= \boldsymbol{Z}^T\left((1+\exp(-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{Z}\boldsymbol{u}))^{-2}\boldsymbol{Z} + \boldsymbol{G}^{-1} - \boldsymbol{\Sigma}_\gamma\right)$$

$$= \boldsymbol{Z}^T\boldsymbol{\pi}(\boldsymbol{1}-\boldsymbol{\pi})\boldsymbol{Z} + \boldsymbol{G}^{-1} - \boldsymbol{\Sigma}_\gamma$$

$$= \boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z} + \boldsymbol{G}^{-1} - \boldsymbol{\Sigma}_\gamma,$$

where $\boldsymbol{\Sigma}_\gamma = \text{diag}\left(p'_\gamma(|\boldsymbol{u}|)(|\boldsymbol{u}|)^{-1}\right)$, is a $m \times m$ diagonal matrix, $p'_\gamma(|\boldsymbol{u}|) = \gamma\,(|\boldsymbol{u}|)\boldsymbol{u}^{-1}$ for LASSO penalty, and $\boldsymbol{W}$ is a $N \times N$ diagonal matrix with the diagonal elements of each block being $\boldsymbol{\pi}_i(\boldsymbol{1}-\boldsymbol{\pi}_i)$ for area i. In order to avoid numerical complications, and to assert the existence of $\boldsymbol{\Sigma}_\gamma$, a non-negative small value $\epsilon(\approx 10^{-10})$ will be introduced to the expression of $\boldsymbol{\Sigma}_\gamma$, $\boldsymbol{\Sigma}_{(\gamma,\epsilon)} = \text{diag}\left(p'_\gamma(|\boldsymbol{u}|)(|\boldsymbol{u}|+\epsilon)^{-1}\right)$ (Lee & Oh, 2009).

Estimation of dispersion parameters

$$h_A = h|_{\hat{\boldsymbol{\tau}}=(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{u}})} + \frac{1}{2}\log\{\det(2\pi\boldsymbol{\mathcal{J}}^{-1})\}|_{\boldsymbol{\tau}=\hat{\boldsymbol{\tau}}}$$

where

$$h|_{\hat{\boldsymbol{\tau}}=(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{u}})} = \ell_1|_{\hat{\boldsymbol{\tau}}=(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{u}})} + \ell_2|_{\hat{\boldsymbol{\tau}}=(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{u}})} - (\gamma|\boldsymbol{u}|)|_{\hat{\boldsymbol{\tau}}=(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{u}})}$$

$$\frac{\partial h_A}{\partial \sigma^2} = \left.\frac{\partial h}{\partial \sigma^2}\right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}},\boldsymbol{u}=\hat{\boldsymbol{u}}} - \frac{1}{2}\text{trace}\left(\boldsymbol{\mathcal{J}}^{-1}\frac{\partial \boldsymbol{\mathcal{J}}}{\partial \sigma^2}\right)\Bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}},\boldsymbol{u}=\hat{\boldsymbol{u}}}$$

$$= \left(-\frac{m}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{m}\hat{u}_i^2 \right)\Bigg|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}},\boldsymbol{u}=\widehat{\boldsymbol{u}}}$$

$$-\frac{1}{2}\mathrm{trace}\left(\begin{pmatrix} \boldsymbol{\mathcal{J}}_{11} & \boldsymbol{\mathcal{J}}_{12} \\ \boldsymbol{\mathcal{J}}_{21} & \boldsymbol{\mathcal{J}}_{22} \end{pmatrix}^{-1} \frac{\partial}{\partial\sigma^2} \begin{pmatrix} \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{Z} \\ \boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{X} & \boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z} + (\sigma^2)^{-1} - \boldsymbol{\Sigma}_\gamma \end{pmatrix} \right)\Bigg|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}},\boldsymbol{u}=\widehat{\boldsymbol{u}}}$$

$$= \left(-\frac{m}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{m}\hat{u}_i^2 \right)\Bigg|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}},\boldsymbol{u}=\widehat{\boldsymbol{u}}} - \frac{1}{2}\mathrm{trace}\left(\begin{pmatrix} \boldsymbol{\mathcal{J}}_{11}^p & \boldsymbol{\mathcal{J}}_{12}^p \\ \boldsymbol{\mathcal{J}}_{21}^p & \boldsymbol{\mathcal{J}}_{22}^p \end{pmatrix} \begin{pmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & -\sigma^{-4}I_{m\times m} \end{pmatrix} \right)\Bigg|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}},\boldsymbol{u}=\widehat{\boldsymbol{u}}}$$

$$= \left(-\frac{m}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{m}\hat{u}_i^2 \right)\Bigg|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}},\boldsymbol{u}=\widehat{\boldsymbol{u}}} + \frac{1}{2\sigma^4}\mathrm{trace}\left(\boldsymbol{\mathcal{J}}_{22}^p\right)\Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}},\boldsymbol{u}=\widehat{\boldsymbol{u}}},$$

Set $\partial h_A/\partial\sigma^2 = 0$,

$$\widehat{\sigma^2} = \frac{1}{m}\left(\sum_{i=1}^{m}\hat{u}_i^2\right)\Bigg|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}},\boldsymbol{u}=\widehat{\boldsymbol{u}}} + \frac{1}{m}\mathrm{trace}\left(\boldsymbol{\mathcal{J}}_{22}^p\right)\Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}},\boldsymbol{u}=\widehat{\boldsymbol{u}}}$$

where $\boldsymbol{\mathcal{J}}_{22}^p = (-\partial^2 h|\partial\boldsymbol{u}^2|_{\boldsymbol{u}=\widehat{\boldsymbol{u}}})^{-1} = \left(\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z} + (\sigma^2)^{-1} - \boldsymbol{\Sigma}_\gamma\right)^{-1}$.

The tuning parameter γ can be obtained minimizing the cross validation (CV), generalized cross validation (GCV), or the Bayesian Information Criterion (BIC) (Fan & Li, 2002, 2006; Ha, Pan, Oh, & Lee, 2014; Lee & Oh, 2009). The use of BIC method is proved that it selects tuning parameters with negligible overfitting error compared to CV or GCV method. Hence, we use BIC to estimate tuning parameters γ,

$$BIC = -2h_A|_{(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{u}})} + e(\gamma)\log N,$$

where $e(\gamma) = trace\left[\left(\left(\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z} + \boldsymbol{G}^{-1} - \boldsymbol{\Sigma}_\gamma\right)^{-1}\left(\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z} + \boldsymbol{G}^{-1} - \boldsymbol{\Sigma}_\gamma\right)\right)\Big|_{(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{u}})}\right]$ is the effective number of parameters in the model,

$$h_A = h|_{(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{u}})} + \frac{1}{2}\log\{\det(2\pi\boldsymbol{\mathcal{J}}^{-1})\}|_{(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{u}})},$$

$$h|_{(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{u}})} = \ell_1|_{(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{u}})} + \ell_2|_{(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{u}})} - (\gamma|\boldsymbol{u}|)|_{(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{u}})}.$$

$$\ell_1 = \boldsymbol{y}^T(\boldsymbol{X\beta} + \boldsymbol{Zu}) - \mathbf{1}^T\log(1 + \exp(\boldsymbol{X\beta} + \boldsymbol{Zu}))$$

$$\ell_2 = -\frac{m}{2}\log 2\pi - \frac{m}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\boldsymbol{u}^T\boldsymbol{u}$$

Chapter 6. Joint modeling of multiple outcomes in small area estimation

The joint modeling approach is generally used with longitudinal or repeated measures data analysis where the measurements are taken over time and time is considered as the random effects component. Here, a similar idea is applied in small area estimation where small areas (clusters) are considered as the random effects component part. It is often possible to observe multiple outcomes from the same individual, hence they might be associated (Ha, Noh, & Lee, 2017; Lee, Ronnegard, & Noh, 2017; Tsiatis & Davidian, 2004). Building separate models without considering this association might not provide accurate results or might lose some significant information. To the best of our knowledge, the joint modelling approach had not been considered in SAE to deal with such association. Therefore, in this chapter, we will consider joint analysis using a joint modeling approach to account for the association of multiple outcomes that could occur from the same subject. Those outcomes are joint through unobserved area-specific random effects which explain the association between multiple outcomes, so it will ignore the biased results which could occur by conducting a separate analysis of multiple outcomes.

Suppose $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_r$ be a vector of r outcomes of interest which are measured on the number of individuals in small areas. Conducting joint analysis on multiple outcomes prevails upon separate analysis to study the association among them. Consider the joint model in SAE for $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_r$ outcomes

$$\begin{pmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \\ \vdots \\ \boldsymbol{Y}_r \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}\boldsymbol{\beta}^{(1)} + \rho^{(1)}\boldsymbol{Z}^{(1)}\boldsymbol{u} \\ \boldsymbol{X}\boldsymbol{\beta}^{(2)} + \rho^{(2)}\boldsymbol{Z}^{(2)}\boldsymbol{u} \\ \vdots \\ \boldsymbol{X}\boldsymbol{\beta}^{(r)} + \rho^{(r)}\boldsymbol{Z}^{(r)}\boldsymbol{u} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\epsilon}_1 \\ \boldsymbol{\epsilon}_2 \\ \vdots \\ \boldsymbol{\epsilon}_r \end{pmatrix}, \tag{6.1}$$

where $\rho^{(1)}, \rho^{(2)}, \dots, \rho^{(r)}$ are shared parameters between outcomes, $\boldsymbol{X}$ is covariate information, $\boldsymbol{\beta}^{(1)}, \dots, \boldsymbol{\beta}^{(r)}$ are fixed effects coefficients, and $\boldsymbol{\beta}^{(.)} = (\beta_1, \dots, \beta_p), p$ is the number of fixed effects

for each outcome, $\boldsymbol{Z}^{(1)}, \ldots, \boldsymbol{Z}^{(r)}$ are diagonal matrices with diagonal elements being 1, and $\boldsymbol{u} = (u_1, \ldots, u_m)$ is a vector of random effects, m is the number of small areas.

Now, we define the joint density for multiple outcomes, $(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_r)$

$$f(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_r) = f(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_r|\boldsymbol{u}) \times f(\boldsymbol{u})$$
$$= f(\boldsymbol{Y}_1|\boldsymbol{u}) \times f(\boldsymbol{Y}_2|\boldsymbol{u}) \times \ldots \times f(\boldsymbol{Y}_r|\boldsymbol{u}) \times f(\boldsymbol{u})$$

Here, we assume that the outcomes are conditionally independent, that is $\boldsymbol{Y}_1|\boldsymbol{u}, \ldots, \boldsymbol{Y}_r|\boldsymbol{u}$ are independent. Now, h-likelihood can be written as

$$h = \ell(\boldsymbol{Y}_1|\boldsymbol{u}) + \cdots + \ell(\boldsymbol{Y}_r|\boldsymbol{u}) + l(\boldsymbol{u}).$$

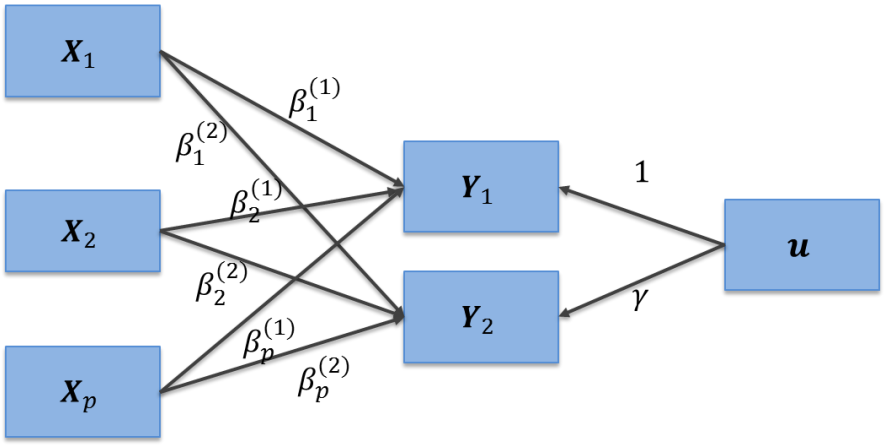


Figure 6.1: Joint modelling of multiple outcomes

In this article, we consider a joint model with two binary response variables, $y = (y_{1ij}, y_{2ij})', i = 1, \ldots, n_j, j = 1, \ldots, m$ where, $y_{1ij}|u_{1i} \sim Bernoulli(p_1), y_{2ij}|u_{2i} \sim Bernoulli(p_2)$, and $u_{1i}, u_{2i} \sim N(0, \theta)$. The auxiliary information

$$X = \begin{pmatrix} \boldsymbol{x}_{1ij}{}' & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{x}_{2ij}{}' \end{pmatrix},$$

where $\boldsymbol{x}'_{1ij} = (x_{1ij1}, \ldots, x_{1ijp_1})'$, and $\boldsymbol{x}'_{2ij} = (x_{2ij1}, \ldots, x_{2ijp_2})'$ are vectors of p_1 and p_2 covariates of y_{1ij} and y_{2ij} respectively. The logit model for y_{1ij} and y_{2ij}

$$P(y_{1ij} = 1|u_{1i}) = \text{logit}^{-1}(\boldsymbol{x}_{1ij}^t\boldsymbol{\beta} + u_{1i}) = \text{logit}^{-1}(\boldsymbol{x}_{1ij}^t\boldsymbol{\beta} + u_i),$$

$$P(y_{2ij} = 1|u_{2i}) = \text{logit}^{-1}(\boldsymbol{x}_{2ij}^{t}\boldsymbol{\delta} + u_{2i}) = \text{logit}^{-1}(\boldsymbol{x}_{2ij}^{t}\boldsymbol{\delta} + \gamma u_i),$$

where $y_r = (y_{rij}, 1 \le i \le m_r, 1 \le j \le n_i), r = 1, 2$ is the binary response vectors for 1st and 2nd outcomes of interest, which are independent given the random effects $u_1, \dots, u_{m_r}$, and γ is the shared parameter. Let $\boldsymbol{u} = (u_i)_{1\le i\le m_1}$, and $\boldsymbol{\beta} = (\beta_k)_{1\le k\le p_1}$ and $\boldsymbol{\delta} = (\delta_k)_{1\le k\le p_2}$ are vectors of unknown fixed effects of y_1 and y_2. Consider the joint h-likelihood in matrix notation

$$h = \ell_{y_1} + \ell_{y_2} + \ell_u$$

$$= \boldsymbol{y}_1^T(\boldsymbol{X_\beta}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}) - \mathbf{1}^T \log(1 + \exp(\boldsymbol{X_\beta}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u})) + \boldsymbol{y}_2^T(\boldsymbol{X}_\delta\boldsymbol{\delta} + \gamma\boldsymbol{Z}\boldsymbol{u})$$

$$- \mathbf{1}^T \log(1 + \exp(\boldsymbol{X}_\delta\boldsymbol{\delta} + \gamma\boldsymbol{Z}\boldsymbol{u})) - \frac{1}{2}\boldsymbol{u}^T\boldsymbol{\theta}^{-1}\boldsymbol{u} - \frac{1}{2}\log(\det(\boldsymbol{\theta})) + c, \qquad (6.2)$$

where $\boldsymbol{\beta}, \boldsymbol{\delta}$ are vector of fixed effects, $\boldsymbol{u}, \gamma\boldsymbol{u}$ are random effects of response variables $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ respectively, with shared parameter γ. $\boldsymbol{X_\beta}$, $\boldsymbol{X}_\delta$ are corresponding design matrices, and $\boldsymbol{\theta}$ is the variance-covariance matrix of random effect $\boldsymbol{u}$. When $\gamma = 0$, $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ are not associated and when $\gamma > 0$, they are positively associated. The parameters of fixed effects and random effects are estimated using the Newton Raphson approximation taking partial derivative of joint h-likelihood with respect to $\boldsymbol{\beta}, \boldsymbol{\delta}$, and $\boldsymbol{u}$. The dispersion parameters, $\boldsymbol{\theta}$ and γ are estimated using the adjusted profile h-likelihood (Ha et al., 2017; Lee et al., 2017).

6.1 Parameter Estimation of Fixed Effects and Random Effects

The partial derivative of h-likelihood equation (6.2) with respect to $\boldsymbol{\beta}$, $\boldsymbol{\delta}$, and $\boldsymbol{u}$, respectively,

$$\frac{\partial h}{\partial \boldsymbol{\beta}} = \left[\boldsymbol{X_\beta}^T\boldsymbol{y}_1 - \boldsymbol{X_\beta}^T(1 + \exp(\boldsymbol{X_\beta}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}))^{-1}\exp(\boldsymbol{X_\beta}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u})\right]_{p\times 1}$$

$$= \boldsymbol{X_\beta}^T\boldsymbol{y}_1 - \boldsymbol{X_\beta}^T\boldsymbol{\pi}_1 = \boldsymbol{X_\beta}^T(\boldsymbol{y}_1 - \boldsymbol{\pi}_1),$$

$$\frac{\partial h}{\partial \boldsymbol{\delta}} = \left[\boldsymbol{X}_\delta{}^T\boldsymbol{y}_2 - \boldsymbol{X}_\delta{}^T(1 + \exp(\boldsymbol{X}_\delta\boldsymbol{\delta} + \gamma\boldsymbol{Z}\boldsymbol{u}))^{-1}\exp(\boldsymbol{X}_\delta\boldsymbol{\delta} + \gamma\boldsymbol{Z}\boldsymbol{u})\right]_{p\times 1}$$

$$= \boldsymbol{X}_\delta{}^T\boldsymbol{y}_2 - \boldsymbol{X}_\delta{}^T\boldsymbol{\pi}_2 = \boldsymbol{X}_\delta{}^T(\boldsymbol{y}_2 - \boldsymbol{\pi}_2),$$

$$\frac{\partial h}{\partial \boldsymbol{u}} = \boldsymbol{Z}^T\boldsymbol{y}_1 - \boldsymbol{Z}^T\boldsymbol{\pi}_1 + \gamma\boldsymbol{Z}^T\boldsymbol{y}_2 - \gamma\boldsymbol{Z}^T\boldsymbol{\pi}_2 - \boldsymbol{\theta}^{-1}\boldsymbol{u}$$

$$= \boldsymbol{Z}^T(\boldsymbol{y}_1 + \gamma\boldsymbol{y}_2 - (\boldsymbol{\pi}_1 + \gamma\boldsymbol{\pi}_2)) - \boldsymbol{\theta}^{-1}\boldsymbol{u},$$

where $\boldsymbol{\pi}_1 = 1/\left(1 + \exp\left(-\boldsymbol{X}_\beta\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u}\right)\right)$, and $\boldsymbol{\pi}_2 = 1/(1 + \exp(-\boldsymbol{X}_{\boldsymbol{\delta}}\boldsymbol{\delta} - \gamma\boldsymbol{Z}\boldsymbol{u}))$. The score function of joint *h*-likelihood can be written as

$$\boldsymbol{S}(\boldsymbol{\tau}) = \begin{pmatrix} \frac{\partial h}{\partial \boldsymbol{\beta}} \\ \frac{\partial h}{\partial \boldsymbol{\delta}} \\ \frac{\partial h}{\partial \boldsymbol{u}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_\beta{}^T(\boldsymbol{y}_1 - \boldsymbol{\pi}_1) \\ \boldsymbol{X}_{\boldsymbol{\delta}}{}^T(\boldsymbol{y}_2 - \boldsymbol{\pi}_2) \\ \boldsymbol{Z}^T(\boldsymbol{y}_1 + \gamma\boldsymbol{y}_2 - (\boldsymbol{\pi}_1 + \gamma\boldsymbol{\pi}_2)) - \boldsymbol{\theta}^{-1}\boldsymbol{u} \end{pmatrix},$$

where $\boldsymbol{\tau} = (\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{u})$ are vectors for fixed effects coefficients of output variables y_1, y_2, and random effects, respectively.

Now, consider the Fisher information matrix of joint *h*-likelihood

$$\boldsymbol{\mathcal{J}} = \begin{bmatrix} -\frac{\partial^2 h}{\partial \boldsymbol{\beta}^2} & -\frac{\partial^2 h}{\partial \boldsymbol{\beta}\partial \boldsymbol{\delta}} & -\frac{\partial^2 h}{\partial \boldsymbol{\beta}\partial \boldsymbol{u}} \\ -\frac{\partial^2 h}{\partial \boldsymbol{\delta}\partial \boldsymbol{\beta}} & -\frac{\partial^2 h}{\partial \boldsymbol{\delta}^2} & -\frac{\partial^2 h}{\partial \boldsymbol{\delta}\partial \boldsymbol{u}} \\ -\frac{\partial^2 h}{\partial \boldsymbol{u}\partial \boldsymbol{\beta}} & -\frac{\partial^2 h}{\partial \boldsymbol{u}\partial \boldsymbol{\delta}} & -\frac{\partial^2 h}{\partial \boldsymbol{u}^2} \end{bmatrix}.$$

The elements of $\boldsymbol{\mathcal{J}}$ are obtained as in chapter 4.3 taking the partial derivative of score function $\boldsymbol{S}(\boldsymbol{\tau})$ with respect to $\boldsymbol{\tau} = (\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{u})$. Taking partial derivatives of $\boldsymbol{\pi}_1$, and $\boldsymbol{\pi}_2$ with respect to $\boldsymbol{\beta}, \boldsymbol{\delta}$ and $\boldsymbol{u}$,

$$\frac{\partial \boldsymbol{\pi}_1}{\partial \boldsymbol{u}} = -\left(1 + \exp\left(-\boldsymbol{X}_\beta\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u}\right)\right)^{-1}\exp\left(-\boldsymbol{X}_\beta\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u}\right)\boldsymbol{Z} = -\boldsymbol{\pi}_1(\boldsymbol{1} - \boldsymbol{\pi}_1)\boldsymbol{Z},$$

$$\frac{\partial \boldsymbol{\pi}_1}{\partial \boldsymbol{\beta}} = -\left(1 + \exp\left(-\boldsymbol{X}_\beta\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u}\right)\right)^{-1}\exp\left(-\boldsymbol{X}_\beta\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u}\right)\boldsymbol{X}_\beta = -\boldsymbol{\pi}_1(\boldsymbol{1} - \boldsymbol{\pi}_1)\boldsymbol{X}_\beta,$$

$$\frac{\partial \boldsymbol{\pi}_2}{\partial \boldsymbol{u}} = -\gamma(1 + \exp(-\boldsymbol{X}_{\boldsymbol{\delta}}\boldsymbol{\delta} - \gamma\boldsymbol{Z}\boldsymbol{u}))^{-1}\exp(-\boldsymbol{X}_{\boldsymbol{\delta}}\boldsymbol{\delta} - \gamma\boldsymbol{Z}\boldsymbol{u}) = -\gamma\boldsymbol{\pi}_2(\boldsymbol{1} - \boldsymbol{\pi}_2)\boldsymbol{Z},$$

$$\frac{\partial \boldsymbol{\pi}_2}{\partial \boldsymbol{\delta}} = -(1 + \exp(-\boldsymbol{X}_{\boldsymbol{\delta}}\boldsymbol{\delta} - \gamma\boldsymbol{Z}\boldsymbol{u}))^{-1}\exp(-\boldsymbol{X}_{\boldsymbol{\delta}}\boldsymbol{\delta} - \gamma\boldsymbol{Z}\boldsymbol{u})\,\boldsymbol{X}_{\boldsymbol{\delta}} = -\boldsymbol{\pi}_2(\boldsymbol{1} - \boldsymbol{\pi}_2)\boldsymbol{X}_{\boldsymbol{\delta}}.$$

The diagonal elements of the matrix $\boldsymbol{\mathcal{J}}$,

$$-\frac{\partial^2 h}{\partial \boldsymbol{\beta}^2} = \boldsymbol{X_\beta}^T\left(1+\exp\left(-\boldsymbol{X_\beta\beta}-\boldsymbol{Zu}\right)\right)^{-1}\exp\left(-\boldsymbol{X_\beta\beta}-\boldsymbol{Zu}\right)\boldsymbol{X_\beta}$$

$$= \boldsymbol{X_\beta}^T\boldsymbol{\pi}_1(\mathbf{1}-\boldsymbol{\pi}_1)\boldsymbol{X_\beta}$$

$$= \boldsymbol{X_\beta}^T\boldsymbol{W}_1\boldsymbol{X_\beta},$$

$$-\frac{\partial^2 h}{\partial \boldsymbol{\delta}^2} = \boldsymbol{X_\delta}^T(1+\exp(-\boldsymbol{X_\delta\delta}-\gamma\boldsymbol{Zu}))^{-1}\exp(-\boldsymbol{X_\delta\delta}-\gamma\boldsymbol{Zu})\,\boldsymbol{X_\delta}$$

$$= \boldsymbol{X_\delta}^T\boldsymbol{\pi}_2(\mathbf{1}-\boldsymbol{\pi}_2)\boldsymbol{X_\delta}$$

$$= \boldsymbol{X_\delta}^T\boldsymbol{W}_2\boldsymbol{X_\delta},$$

$$-\frac{\partial^2 h}{\partial \boldsymbol{u}^2} = \boldsymbol{Z}^T\boldsymbol{\pi}_1(\mathbf{1}-\boldsymbol{\pi}_1)\boldsymbol{Z}+\gamma^2\boldsymbol{Z}^T\boldsymbol{\pi}_2(\mathbf{1}-\boldsymbol{\pi}_2)\boldsymbol{Z}+\boldsymbol{\theta}^{-1}$$

$$= \boldsymbol{Z}^T\boldsymbol{W}_1\boldsymbol{Z}+\boldsymbol{Z}^T(\gamma^2\boldsymbol{W}_2)\boldsymbol{Z}+\boldsymbol{\theta}^{-1}.$$

The off-diagonal elements of the matrix $\boldsymbol{\mathcal{J}}$,

$$-\frac{\partial^2 h}{\partial \boldsymbol{\beta}\partial\boldsymbol{\delta}} = -\frac{\partial^2 h}{\partial \boldsymbol{\delta}\partial\boldsymbol{\beta}} = 0,$$

$$-\frac{\partial^2 h}{\partial \boldsymbol{\beta}\partial\boldsymbol{u}} = -\frac{\partial h}{\partial \boldsymbol{u}}\left(\boldsymbol{X_\beta}^T(\boldsymbol{y}_1-\boldsymbol{\pi}_1)\right) = \boldsymbol{X_\beta}^T\boldsymbol{\pi}_1(\mathbf{1}-\boldsymbol{\pi}_1)\boldsymbol{Z} = \boldsymbol{X_\beta}^T\boldsymbol{W}_1\boldsymbol{Z},$$

$$-\frac{\partial^2 h}{\partial \boldsymbol{\delta}\partial\boldsymbol{u}} = -\frac{\partial}{\partial \boldsymbol{u}}\left(\boldsymbol{X_\delta}^T(\boldsymbol{y}_2-\boldsymbol{\pi}_2)\right) = \boldsymbol{X_\delta}^T\gamma\boldsymbol{\pi}_2(\mathbf{1}-\boldsymbol{\pi}_2)\boldsymbol{Z} = \boldsymbol{X_\delta}^T(\gamma\boldsymbol{W}_2)\boldsymbol{Z},$$

$$-\frac{\partial^2 h}{\partial \boldsymbol{u}\partial\boldsymbol{\beta}} = -\frac{\partial}{\partial \boldsymbol{\beta}}\left(\boldsymbol{Z}^T\left(\boldsymbol{y}_1+\gamma\boldsymbol{y}_2-(\boldsymbol{\pi}_1+\gamma\boldsymbol{\pi}_2)\right)-\boldsymbol{\theta}^{-1}\boldsymbol{u}\right)$$

$$= \boldsymbol{Z}^T\boldsymbol{\pi}_1(\mathbf{1}-\boldsymbol{\pi}_1)\boldsymbol{X_\beta} = \boldsymbol{Z}^T\boldsymbol{W}_1\boldsymbol{X_\beta},$$

$$-\frac{\partial^2 h}{\partial \boldsymbol{u}\partial\boldsymbol{\delta}} = -\frac{\partial}{\partial \boldsymbol{\delta}}\left(\boldsymbol{Z}^T\left(\boldsymbol{y}_1+\gamma\boldsymbol{y}_2-(\boldsymbol{\pi}_1+\gamma\boldsymbol{\pi}_2)\right)-\boldsymbol{\theta}^{-1}\boldsymbol{u}\right)$$

$$= \boldsymbol{Z}^T\gamma\boldsymbol{\pi}_2(\mathbf{1}-\boldsymbol{\pi}_2)\boldsymbol{X_\delta} = \boldsymbol{Z}^T(\gamma\boldsymbol{W}_2)\boldsymbol{X_\delta}.$$

Using the above quantities $\boldsymbol{\mathcal{J}}$ can be written as

$$\boldsymbol{\mathcal{J}} = \begin{bmatrix} \boldsymbol{X_\beta}^T\boldsymbol{W}_1\boldsymbol{X_\beta} & \mathbf{0} & \boldsymbol{X_\beta}^T\boldsymbol{W}_1\boldsymbol{Z} \\ \mathbf{0} & \boldsymbol{X_\delta}^T\boldsymbol{W}_2\boldsymbol{X_\delta} & \boldsymbol{X_\delta}^T(\gamma\boldsymbol{W}_2)\boldsymbol{Z} \\ \boldsymbol{Z}^T\boldsymbol{W}_1\boldsymbol{X_\beta} & \boldsymbol{Z}^T(\gamma\boldsymbol{W}_2)\boldsymbol{X_\delta} & \boldsymbol{Z}^T\boldsymbol{W}_1\boldsymbol{Z}+\boldsymbol{Z}^T(\gamma^2\boldsymbol{W}_2)\boldsymbol{Z}+\boldsymbol{\theta}^{-1} \end{bmatrix},$$

where $\boldsymbol{W}_r = \boldsymbol{\pi}_r(\mathbf{1}-\boldsymbol{\pi}_r), r = 1,2.$

Now, the MHLEs are obtained using Newton Raphson approximation,

$$\begin{pmatrix} \widehat{\boldsymbol{\beta}}^{(k+1)} \\ \widehat{\boldsymbol{\delta}}^{(k+1)} \\ \widehat{\boldsymbol{u}}^{(k+1)} \end{pmatrix} = \begin{pmatrix} \widehat{\boldsymbol{\beta}}^{(k)} \\ \widehat{\boldsymbol{\delta}}^{(k)} \\ \widehat{\boldsymbol{u}}^{(k)} \end{pmatrix} + \left(\boldsymbol{\mathcal{J}}^{-1}\boldsymbol{\mathcal{S}}(\tau)\right)|_{\boldsymbol{\tau}=\hat{\boldsymbol{\tau}}^{(k)}}, \tag{6.3}$$

where $\boldsymbol{\tau} = (\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{u})$, and $\hat{\boldsymbol{\tau}} = \left(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{u}}\right)$.

6.2 Parameter Estimation of variance component and shared parameter

The HMLEs of θ and γ are obtained using the adjusted h-likelihood by solving the score equations $\partial h_A/\partial\theta = \partial h_A/\partial\gamma = 0$. First, consider the equation (4.5) to estimate of θ at current estimates $\boldsymbol{\tau} = \hat{\boldsymbol{\tau}}$

$$\frac{\partial h_A}{\partial\theta} = \frac{\partial h}{\partial\theta}\Big|_{\hat{\boldsymbol{\tau}}} - \frac{1}{2}\text{trace}\left(\boldsymbol{\mathcal{J}}^{-1}\frac{\partial\boldsymbol{\mathcal{J}}}{\partial\theta}\right)\Big|_{\hat{\boldsymbol{\tau}}}. \tag{6.4}$$

Note that $\partial\ell_{1ij}/\partial\theta = \partial\ell_{2ij}/\partial\theta = 0$, the score function can be written as

$$\frac{\partial h}{\partial\theta} = 0 + \frac{\partial}{\partial\theta}\left(\sum_{i=1}^{m}\ell_{3i}\right).$$

The adjusted profile log-likelihood of $u_i \sim N(0,\theta)$ for small area i

$$\ell_{2i} = -\frac{m}{2}\log 2\pi - \frac{m}{2}\log\theta - \frac{1}{2\theta}\sum_{i=1}^{m}u_i^2 = -\frac{m}{2}\log 2\pi - \frac{m}{2}\log\theta - \frac{1}{2\theta}\boldsymbol{u}^T\boldsymbol{u},$$

where $\theta I_{m\times m} = \sigma^2 I_{m\times m}$. Now, from $h = \sum_{ij}\ell_{1ij} + \sum_{ij}\ell_{2ij} + \sum_i \ell_{2i}$, $\partial h/\partial\theta$ given $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}, \boldsymbol{u} = \widehat{\boldsymbol{u}}$ can be represented as

$$\frac{\partial h}{\partial\theta}\Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}},\boldsymbol{\delta}=\widehat{\boldsymbol{\delta}},\boldsymbol{u}=\widehat{\boldsymbol{u}}} = 0 + \frac{\partial}{\partial\theta}\left(-\frac{m}{2}\log\theta - \frac{1}{2\theta}\sum_{i=1}^{m}\hat{u}_i^2\right)$$

$$= \left(-\frac{m}{2\theta} + \frac{1}{2\theta^2}\sum_{i=1}^{m}\hat{u}_i^2\right)\Big|_{\hat{\boldsymbol{\tau}}}.$$

The partial derivative of the observed information matrix with respect to θ given $\boldsymbol{\tau} = \hat{\boldsymbol{\tau}}$

$$\frac{\partial\boldsymbol{\mathcal{J}}}{\partial\theta}\Big|_{\hat{\boldsymbol{\tau}}} = \frac{\partial}{\partial\theta}\begin{bmatrix} -\frac{\partial^2 h}{\partial\boldsymbol{\beta}^2} & -\frac{\partial^2 h}{\partial\boldsymbol{\beta}\partial\boldsymbol{\delta}} & -\frac{\partial^2 h}{\partial\boldsymbol{\beta}\partial\boldsymbol{u}} \\ -\frac{\partial^2 h}{\partial\boldsymbol{\delta}\partial\boldsymbol{\beta}} & -\frac{\partial^2 h}{\partial\boldsymbol{\delta}^2} & -\frac{\partial^2 h}{\partial\boldsymbol{\delta}\partial\boldsymbol{u}} \\ -\frac{\partial^2 h}{\partial\boldsymbol{u}\partial\boldsymbol{\beta}} & -\frac{\partial^2 h}{\partial\boldsymbol{u}\partial\boldsymbol{\delta}} & -\frac{\partial^2 h}{\partial\boldsymbol{u}^2} \end{bmatrix},$$

$$= \frac{\partial}{\partial\theta}\begin{bmatrix} \boldsymbol{X_\beta}^T\boldsymbol{W}_1\boldsymbol{X}_\beta & \mathbf{0} & \boldsymbol{X_\beta}^T\boldsymbol{W}_1\boldsymbol{Z} \\ \mathbf{0} & \boldsymbol{X_\delta}^T\boldsymbol{W}_2\boldsymbol{X}_\delta & \boldsymbol{X_\delta}^T(\gamma\boldsymbol{W}_2)\boldsymbol{Z} \\ \boldsymbol{Z}^T\boldsymbol{W}_1\boldsymbol{X}_\beta & \boldsymbol{Z}^T(\gamma\boldsymbol{W}_2)\boldsymbol{X}_\delta & \boldsymbol{Z}^T\boldsymbol{W}_1\boldsymbol{Z} + \boldsymbol{Z}^T(\gamma^2\boldsymbol{W}_2)\boldsymbol{Z} + \boldsymbol{\theta}^{-1} \end{bmatrix},$$

$$= \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & -\theta^{-2}I_{m\times m} \end{bmatrix}.$$

From (6.4)

$$\frac{\partial h_A}{\partial\theta} = \left(-\frac{m}{2\theta} + \frac{1}{2\theta^2}\sum_{i=1}^{m}\hat{u}_i^2\right)\Bigg|_{\hat{\boldsymbol{\tau}}} - \frac{1}{2}\text{trace}\left(\begin{bmatrix} \boldsymbol{\mathcal{J}}_{11}^* & \boldsymbol{\mathcal{J}}_{12}^* & \boldsymbol{\mathcal{J}}_{13}^* \\ \boldsymbol{\mathcal{J}}_{21}^* & \boldsymbol{\mathcal{J}}_{22}^* & \boldsymbol{\mathcal{J}}_{23}^* \\ \boldsymbol{\mathcal{J}}_{31}^* & \boldsymbol{\mathcal{J}}_{32}^* & \boldsymbol{\mathcal{J}}_{33}^* \end{bmatrix}\begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & -\theta^{-2}I_{m\times m} \end{bmatrix}\right)\Bigg|_{\hat{\boldsymbol{\tau}}}$$

$$= \left(-\frac{m}{2\theta} + \frac{1}{2\theta^2}\sum_{i=1}^{m}\hat{u}_i^2\right)\Bigg|_{\hat{\boldsymbol{\tau}}} - \frac{1}{2}\text{trace}\left(\begin{bmatrix} \mathbf{0} & \mathbf{0} & -\boldsymbol{\mathcal{J}}_{13}^*\theta^{-2}I_{m\times m} \\ \mathbf{0} & \mathbf{0} & -\boldsymbol{\mathcal{J}}_{23}^*\theta^{-2}I_{m\times m} \\ \mathbf{0} & \mathbf{0} & -\boldsymbol{\mathcal{J}}_{33}^*\theta^{-2}I_{m\times m} \end{bmatrix}\right)\Bigg|_{\hat{\boldsymbol{\tau}}}$$

$$= \left(-\frac{m}{2\theta} + \frac{1}{2\theta^2}\sum_{i=1}^{m}\hat{u}_i^2\right)\Bigg|_{\hat{\boldsymbol{\tau}}} + \frac{1}{2\theta^2}\text{trace}(\boldsymbol{\mathcal{J}}_{33}^*)\Big|_{\hat{\boldsymbol{\tau}}}$$

Set $\partial h_A/\partial\theta = 0$

$$\left(-\frac{m}{2\theta} + \frac{1}{2\theta^2}\sum_{i=1}^{m}\hat{u}_i^2\right)\Bigg|_{\hat{\boldsymbol{\tau}}} + \frac{1}{2\theta^2}\text{trace}(\boldsymbol{\mathcal{J}}_{33}^*)\Big|_{\hat{\boldsymbol{\tau}}} = 0.$$

Thus,

$$\hat{\theta} = \frac{1}{m}\left(\sum_{i=1}^{m}\hat{u}_i^2\right)\Bigg|_{\hat{\boldsymbol{\tau}}} + \frac{1}{m}\text{trace}(\boldsymbol{\mathcal{J}}_{33}^*)\Big|_{\hat{\boldsymbol{\tau}}}. \tag{6.6}$$

MHLE of θ is obtained using (6.6). Now, the MHLE of γ is obtained using the partial derivative of (4.5) with respect to γ

$$\frac{\partial h_A}{\partial\gamma} = \frac{\partial h}{\partial\gamma}\Big|_{\hat{\boldsymbol{\tau}}} - \frac{1}{2}\text{trace}\left(\boldsymbol{\mathcal{J}}^{-1}\frac{\partial\boldsymbol{\mathcal{J}}}{\partial\gamma}\right)\Big|_{\hat{\boldsymbol{\tau}}}. \tag{6.7}$$

Consider the first term of (6.7)

$$\frac{\partial h}{\partial\gamma}\Big|_{\hat{\boldsymbol{\tau}}} = (\boldsymbol{Zu})^T\boldsymbol{y}_2 - (\boldsymbol{Zu})^T(\mathbf{1} + \exp(\boldsymbol{X}_\delta\boldsymbol{\delta} + \gamma\boldsymbol{Zu}))^{-1}\exp(\boldsymbol{X}_\delta\boldsymbol{\delta} + \gamma\boldsymbol{Zu})$$

$$= ((\boldsymbol{Zu})^T\boldsymbol{y}_2 - (\boldsymbol{Zu})^T\boldsymbol{\pi}_2)|_{\hat{\boldsymbol{\tau}}} = (\boldsymbol{Zu})^T(\boldsymbol{y}_2 - \boldsymbol{\pi}_2)|_{\hat{\boldsymbol{\tau}}}$$

The second term of (6.7)

$$\frac{\partial \boldsymbol{\mathcal{J}}}{\partial \gamma} = \frac{\partial}{\partial \gamma}\begin{bmatrix} \boldsymbol{X_\beta}^T \boldsymbol{W}_1 \boldsymbol{X}_\beta & \mathbf{0} & \boldsymbol{X_\beta}^T \boldsymbol{W}_1 \boldsymbol{Z} \\ \mathbf{0} & \boldsymbol{X_\delta}^T \boldsymbol{W}_2 \boldsymbol{X}_\delta & \gamma \boldsymbol{X_\delta}^T \boldsymbol{W}_2 \boldsymbol{Z} \\ \boldsymbol{Z}^T \boldsymbol{W}_1 \boldsymbol{X}_\beta & \gamma \boldsymbol{Z}^T \boldsymbol{W}_2 \boldsymbol{X}_\delta & \boldsymbol{Z}^T (\boldsymbol{W}_1 + \gamma^2 \boldsymbol{W}_2)\boldsymbol{Z} + \boldsymbol{\theta}^{-1} \end{bmatrix}\Bigg|_{\hat{\boldsymbol{\tau}}}$$

$$= \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{X_\delta}^T \dfrac{\partial \boldsymbol{W}_2}{\partial \gamma} \boldsymbol{X}_\delta & \gamma \boldsymbol{X_\delta}^T \dfrac{\partial \boldsymbol{W}_2}{\partial \gamma} \boldsymbol{Z} + \boldsymbol{X_\delta}^T \boldsymbol{W}_2 \boldsymbol{Z} \\ \mathbf{0} & \gamma \boldsymbol{Z}^T \dfrac{\partial \boldsymbol{W}_2}{\partial \gamma} \boldsymbol{X}_\delta + \boldsymbol{Z}^T \boldsymbol{W}_2 \boldsymbol{X}_\delta & \boldsymbol{Z}^T \left(\gamma^2 \dfrac{\partial \boldsymbol{W}_2}{\partial \gamma} + 2\gamma \boldsymbol{W}_2 \right) \boldsymbol{Z} \end{bmatrix}\Bigg|_{\hat{\boldsymbol{\tau}}},$$

where,

$$\frac{\partial \boldsymbol{\pi}_2}{\partial \gamma} = -(1 + \exp(-\boldsymbol{X_\delta}\boldsymbol{\delta} - \gamma \boldsymbol{Z}\boldsymbol{u}))^{-1} \exp(-\boldsymbol{X_\delta}\boldsymbol{\delta} - \gamma \boldsymbol{Z}\boldsymbol{u})\, \boldsymbol{Z}\boldsymbol{u} = -\boldsymbol{\pi}_2(\mathbf{1} - \boldsymbol{\pi}_2)\boldsymbol{Z}\boldsymbol{u}$$

then,

$$\frac{\partial \boldsymbol{W}_2}{\partial \gamma} = \frac{\partial}{\partial \gamma}\left(\boldsymbol{\pi}_2(\mathbf{1} - \boldsymbol{\pi}_2)\right) = -\boldsymbol{\pi}_2 \frac{\partial \boldsymbol{\pi}_2}{\partial \gamma} + (\mathbf{1} - \boldsymbol{\pi}_2)\frac{\partial \boldsymbol{\pi}_2}{\partial \gamma}$$

$$= \boldsymbol{\pi}_2(\mathbf{1} - \boldsymbol{\pi}_2)(2\boldsymbol{\pi}_2 - \mathbf{1})\boldsymbol{Z}\boldsymbol{u}.$$

Now, we take the partial derivative of the (6.7) with respect to γ to obtain the Hessian matrix $\boldsymbol{\mathcal{J}}_A$

$$\boldsymbol{\mathcal{J}}_A = \frac{\partial^2 h_A}{\partial \gamma^2} = \frac{\partial^2 h}{\partial \gamma^2}\Big|_{\hat{\boldsymbol{\tau}}} - \frac{1}{2}\frac{\partial}{\partial \gamma}\left(\operatorname{trace}\left(\boldsymbol{\mathcal{J}}^{-1}\frac{\partial \boldsymbol{\mathcal{J}}}{\partial \gamma}\right)\right)\Bigg|_{\hat{\boldsymbol{\tau}}}. \qquad (6.8)$$

The expressions (6.7) and (6.8) will be used in the Newton Raphson procedure to obtain MLE of γ

$$\frac{\partial^2 h}{\partial \gamma^2}\Big|_{\hat{\boldsymbol{\tau}}} = \frac{\partial}{\partial \gamma}(\boldsymbol{Z}\boldsymbol{u})^T(\boldsymbol{y}_2 - \boldsymbol{\pi}_2)\Big|_{\hat{\boldsymbol{\tau}}} = -(\boldsymbol{Z}\boldsymbol{u})^T \frac{\partial}{\partial \gamma}\boldsymbol{\pi}_2\Big|_{\hat{\boldsymbol{\tau}}}$$

$$\frac{\partial^2 h}{\partial \gamma^2}\Big|_{\hat{\boldsymbol{\tau}}} = (\boldsymbol{Z}\boldsymbol{u})^T \boldsymbol{\pi}_2(\mathbf{1} - \boldsymbol{\pi}_2)\boldsymbol{Z}\boldsymbol{u}|_{\hat{\boldsymbol{\tau}}}. \qquad (6.9)$$

The second term of (6.8)

$$\frac{\partial}{\partial \gamma}\left(\operatorname{trace}\left(\boldsymbol{\mathcal{J}}^{-1}\frac{\partial \boldsymbol{\mathcal{J}}}{\partial \gamma}\right)\right)\Bigg|_{\hat{\boldsymbol{\tau}}} =$$

$$\frac{\partial}{\partial \gamma}\left(\operatorname{trace}\left(\begin{bmatrix} \boldsymbol{\mathcal{J}}^*_{11} & \boldsymbol{\mathcal{J}}^*_{12} & \boldsymbol{\mathcal{J}}^*_{13} \\ \boldsymbol{\mathcal{J}}^*_{21} & \boldsymbol{\mathcal{J}}^*_{22} & \boldsymbol{\mathcal{J}}^*_{23} \\ \boldsymbol{\mathcal{J}}^*_{31} & \boldsymbol{\mathcal{J}}^*_{32} & \boldsymbol{\mathcal{J}}^*_{33} \end{bmatrix}\begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{X_\delta}^T \dfrac{\partial \boldsymbol{W}_2}{\partial \gamma} \boldsymbol{X}_\delta & \gamma \boldsymbol{X_\delta}^T \dfrac{\partial \boldsymbol{W}_2}{\partial \gamma} \boldsymbol{Z} + \boldsymbol{X_\delta}^T \boldsymbol{W}_2 \boldsymbol{Z} \\ \mathbf{0} & \gamma \boldsymbol{Z}^T \dfrac{\partial \boldsymbol{W}_2}{\partial \gamma} \boldsymbol{X}_\delta + \boldsymbol{Z}^T \boldsymbol{W}_2 \boldsymbol{X}_\delta & \boldsymbol{Z}^T \left(\gamma^2 \dfrac{\partial \boldsymbol{W}_2}{\partial \gamma} + 2\gamma \boldsymbol{W}_2 \right) \boldsymbol{Z} \end{bmatrix}\right)\right)\Bigg|_{\hat{\boldsymbol{\tau}}} \qquad (6.10)$$

The MLE of θ is obtained using (6.6) and the shared parameter (γ) is obtained using Newton Raphson approximation.

Chapter 7. Simulation

7.1 Simulation – Binary HGLM

7.1.1 Data Generation

The proposed h-likelihood approach is evaluated through Monte Carlo simulation performing 1000 simulations to estimate fixed effects and random effects. The first simulation study was performed to evaluate the proposed method based on a single outcome variable. We considered GLMM as our benchmark method to compare the results of the proposed method varying small areas $(q = 5, 10, 20, 30)$, each small area with the sample sizes of $n_q = 10, 30, 50, 100, 500$, repectively. Total of 20 data sets, the smallest data set with 50 observations and the largest data set with 15000 observations, were analyzed based on two discrete variables with one binary outcome variable. First, consider the binary HGLM model with $y|u \sim Bino(p)$, $u \sim N(0, \sigma^2)$ with log likelihood of $\boldsymbol{y}|\boldsymbol{u}$

$$\boldsymbol{\ell}_{\boldsymbol{y}|\boldsymbol{u}} = \boldsymbol{y}(\boldsymbol{X\beta} + \boldsymbol{Zu}) - \log(1 + \exp(\boldsymbol{X\beta} + \boldsymbol{Zu})),$$

$$\theta = \boldsymbol{X\beta} + \boldsymbol{Zu},$$

$$b(\theta) = \log(1 + \exp \boldsymbol{X\beta} + \boldsymbol{Zu}),$$

$$\phi = 1, u \sim N(0, \sigma^2) \text{ with } \lambda = \sigma^2.$$

The initial random effects for each small area, are simulated from normal distribution with mean 0 and initial variance is 0.1 $(\sigma_0^2 = 0.1)$. The initial values of fixed effects are assumed to be $\beta_{11} = 1.3$, $\beta_{21} = 1.5$, for two discrete variables X_1, and X_2, and the intercept term, $\beta_0 = -1.5$. The probability of $\boldsymbol{y}$ =1 for the binary response variable was calculated using logit model as

$$P_0(\boldsymbol{y} = 1) = \frac{\exp(\boldsymbol{X\beta} + \boldsymbol{Zu})}{1 + \exp(\boldsymbol{X\beta} + \boldsymbol{Zu})},$$

where $\boldsymbol{X} = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_{11} \\ \beta_{12} \\ \beta_{21} \\ \beta_{22} \end{pmatrix} = \begin{pmatrix} -1.5 \\ 1.3 \\ 0.0 \\ 1.5 \\ 0.0 \end{pmatrix}, \boldsymbol{Z} = \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{pmatrix}_{N \times m}, \boldsymbol{u} = \begin{pmatrix} u_1 \\ \vdots \\ u_m \end{pmatrix}$, $\boldsymbol{u}$ is drawn from $N(0, 0.1)$**.** Under this scenario, the binary response variable is simulated using the calculated probability. Now we apply the proposed h-likelihood iterative method to estimate maximum likelihood estimates of $\boldsymbol{\beta}, \boldsymbol{u},$ and σ^2 for the first combination of 20 combinations of (q, n_q) mentioned above. 1000 simulated data sets for each combination of q and n_q were generated, which means 16000 data sets were used to evaluate the proposed method.

7.1.2 MC Simulation Results

As described in section 7.1.1, the proposed h-likelihood approach and GLMM method were applied to 20 combinations of data sets each combination were applied to 1000 different data sets for same q and n_q. The final estimates for each scenario of (q, n_q) were obtained averaging over 1000 simulations for MHLEs and MLEs from GLMM. The results are shown in Table 7.1.

The final MHLEs of $\boldsymbol{\beta}$, and θ are obtained by taking the average over the number of simulations (1000 simulations). The MHLE of $\boldsymbol{u}$ is obtained averaging over 1000 simulations and averaging over sample sizes of each scenario of (q, n_q).

$$\hat{\boldsymbol{\beta}}_{mean} = \frac{1}{s}\sum_{i=1}^{s} \hat{\boldsymbol{\beta}},$$

where s is the total number of simulations ($s = 1000$) performed at each combination of q and n_q. The performance of MHLEs $\hat{\boldsymbol{\psi}} = (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{u}}, \hat{\theta})$ are evaluated using the mean squared error (MSE) and the relative bias

$$Bias(\hat{\boldsymbol{\psi}}) = \frac{E(\hat{\boldsymbol{\psi}}) - \boldsymbol{\psi}}{\boldsymbol{\psi}}.$$

The MSE of $\widehat{\boldsymbol{\beta}}$, and $\hat{\theta}$ are obtained averaging over 1000 simulations for each scenario and the MSE for each small area ($MSE(\widehat{\boldsymbol{u}})$) is obtained averaging over 1000 simulations and number of small areas (q).

$$MSE\left(\widehat{\boldsymbol{\beta}}_{q,n_q,k}\right) = \frac{1}{1000}\sum_{s=1}^{1000}\left(\hat{\beta}_{sk} - \beta_k\right)^2, k = 1, \dots, (p+1).$$

$$MSE\left(\widehat{\boldsymbol{u}}_{q,n_q}\right) = \frac{1}{1000}\sum_{s=1}^{1000}\frac{1}{q}\left(\sum_{i=1}^{q}(\hat{u}_{si} - u_i)^2\right)$$

The simulation results imply that the proposed method obtains reliably better estimates regardless of the sample size of each small area. The HMLE estimate of variance parameter $(\hat{\theta})$ is more accurate than GLMM estimates in every scenario. The variance parameter is underestimated by both methods, but it is reliable. However, the both models, proposed HGLM and GLMM model performs better when the sample size increases. Furthermore, it is promising that the average mean squared error decreases when the sample size increases. Table 7.1 shows the fixed effects estimates, variance component estimates, and root mean squared error (RMSE) for both methods for each combination of $(q, n_q = n)$.

Table 7.1. Benchmark analysis results to compare SAE results of HGLM and GLMM using Monte Carlo (MC) simulation

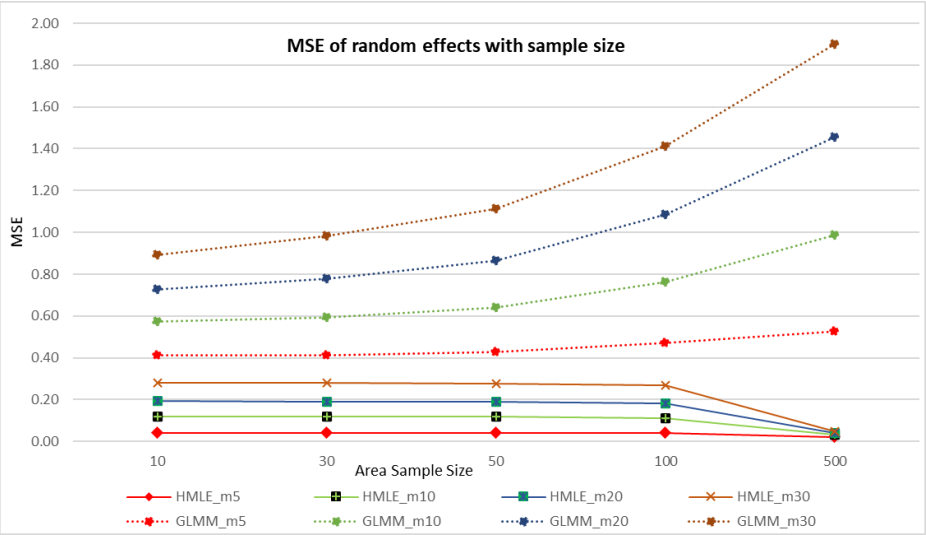
CPH with Bias Correction Estimates					**GLMM Estimates**				
Sample size	**Parameter**	**MHLE**	**RMSE**	**Bias**	**Sample size**	**Parameter**	**MLE**	**RMSE**	**Bias**
$m = 5$	β_0	-1.6627	1.8970	0.1085	$m = 5$	β_0	-1.6761	1.2982	0.1174
$n = 10$	β_{11}	1.5077	1.9148	0.1598	$n = 10$	β_{11}	1.5164	1.3203	0.1664
	β_{21}	1.6968	1.9123	0.1312		β_{21}	1.7139	1.3247	0.1426
	θ	0.0026	0.0974	-0.9736		θ	0.2123	0.3737	1.1232
$m = 5$	β_0	-1.4412	0.3455	-0.0392	$m = 5$	β_0	-1.4575	0.3498	-0.0283
$n = 30$	β_{11}	1.3148	0.3859	0.0113	$n = 30$	β_{11}	1.3303	0.3940	0.0233
	β_{21}	1.5011	0.3823	0.0008		β_{21}	1.5182	0.3888	0.0121
	θ	0.0021	0.0986	-0.9790		θ	0.1333	0.1926	0.3326
$m = 5$	β_0	-1.4518	0.2700	-0.0321	$m = 5$	β_0	-1.4635	0.2703	-0.0243

$n = 50$	β_{11}	1.3055	0.2925	0.0042	$n = 50$	β_{11}	1.3159	0.2945	0.0122
	β_{21}	1.5136	0.2922	0.0091		β_{21}	1.5266	0.2955	0.0177
	θ	0.0017	0.0986	-0.9828		θ	0.1327	0.1598	0.3272
$m = 5$	β_0	-1.4435	0.2012	-0.0376	$m = 5$	β_0	-1.4538	0.1998	-0.0308
$n = 100$	β_{11}	1.2956	0.2008	-0.0034	$n = 100$	β_{11}	1.3058	0.2019	0.0045
	β_{21}	1.5011	0.2115	0.0008		β_{21}	1.5118	0.2139	0.0079
	θ	0.0025	0.0983	-0.9754		θ	0.1421	0.1322	0.4208
$m = 5$	β_0	-1.4427	0.0996	-0.0382	$m = 5$	β_0	-1.4474	0.0972	-0.0351
$n = 500$	β_{11}	1.2997	0.0894	-0.0002	$n = 500$	β_{11}	1.3054	0.0899	0.0042
	β_{21}	1.4955	0.0887	-0.0030		β_{21}	1.5020	0.0889	0.0014
	θ	0.0158	0.0859	-0.8417		θ	0.1874	0.1005	0.8737
$m = 10$	β_0	-1.5054	1.8970	0.0036	$m = 10$	β_0	-1.5481	1.2982	0.0320
$n = 10$	β_{11}	1.3497	1.9148	0.0382	$n = 10$	β_{11}	1.3877	1.3203	0.0674
	β_{21}	1.5284	1.9123	0.0189		β_{21}	1.5723	1.3247	0.0482
	θ	0.0026	0.0974	-0.9736		θ	0.2354	0.3737	1.3543
$m = 10$	β_0	-1.4687	0.4768	-0.0209	$m = 10$	β_0	-1.4924	0.4972	-0.0050
$n = 30$	β_{11}	1.2929	0.5036	-0.0055	$n = 10$	β_{11}	1.3137	0.5257	0.0105
	β_{21}	1.4923	0.4777	-0.0051		β_{21}	1.5176	0.5035	0.0117
	θ	0.0016	0.0974	-0.9839		θ	0.2119	0.3299	1.1190
$m = 10$	β_0	-1.4477	0.2564	-0.0349	$m = 10$	β_0	-1.4697	0.2585	-0.0202
$n = 50$	β_{11}	1.2856	0.2793	-0.0111	$n = 30$	β_{11}	1.3056	0.2851	0.0043
	β_{21}	1.4810	0.2625	-0.0127		β_{21}	1.5038	0.2687	0.0025
	θ	0.0014	0.0984	-0.9864		θ	0.2251	0.2185	1.2506
$m = 10$	β_0	-1.4483	0.1951	-0.0344	$m = 10$	β_0	-1.4703	0.1931	-0.0198
$n = 100$	β_{11}	1.2824	0.2003	-0.0135	$n = 50$	β_{11}	1.3022	0.2033	0.0017
	β_{21}	1.4811	0.2105	-0.0126		β_{21}	1.5045	0.2129	0.0030
	θ	0.0031	0.0986	-0.9692		θ	0.2544	0.1916	1.5435
$m = 10$	β_0	-1.4676	0.1417	-0.0216	$m = 10$	β_0	-1.4691	0.1372	-0.0206
$n = 500$	β_{11}	1.2995	0.1374	-0.0004	$n = 100$	β_{11}	1.3039	0.1388	0.0030
	β_{21}	1.4950	0.1403	-0.0034		β_{21}	1.5002	0.1408	0.0001
	θ	0.0542	0.0975	-0.4580		θ	0.2744	0.1809	1.7437
$m = 20$	β_0	-1.4837	0.3054	-0.0108	$m = 20$	β_0	-1.5184	0.3172	0.0123
$n = 10$	β_{11}	1.2987	0.3310	-0.0010	$n = 10$	β_{11}	1.3293	0.3439	0.0225
	β_{21}	1.4940	0.3210	-0.0040		β_{21}	1.5298	0.3343	0.0199
	θ	0.0026	0.0974	-0.9738		θ	0.2393	0.2797	1.3931
$m = 20$	β_0	-1.4688	0.1722	-0.0208	$m = 20$	β_0	-1.4907	0.1723	-0.0062
$n = 30$	β_{11}	1.2772	0.1860	-0.0175	$n = 30$	β_{11}	1.2963	0.1870	-0.0028
	β_{21}	1.4790	0.1862	-0.0140		β_{21}	1.5017	0.1883	0.0012
	θ	0.0016	0.0984	-0.9841		θ	0.2270	0.1932	1.2704
$m = 20$	β_0	-1.4796	0.1348	-0.0136	$m = 20$	β_0	-1.5008	0.1348	0.0006
$n = 50$	β_{11}	1.2876	0.1473	-0.0096	$n = 50$	β_{11}	1.3059	0.1481	0.0045
	β_{21}	1.4878	0.1414	-0.0081		β_{21}	1.5098	0.1432	0.0066
	θ	0.0013	0.0987	-0.9869		θ	0.2412	0.1762	1.4121
$m = 20$	β_0	-1.4679	0.0993	-0.0214	$m = 20$	β_0	-1.4911	0.0964	-0.0059
$n = 100$	β_{11}	1.2817	0.1019	-0.0141	$n = 100$	β_{11}	1.3022	0.1016	0.0017
	β_{21}	1.4771	0.1031	-0.0153		β_{21}	1.5004	0.1025	0.0003
	θ	0.0014	0.0986	-0.9855		θ	0.2661	0.1772	1.6614

$m=20$	β_0	-1.4911	0.0441	-0.0059	$m=20$	β_0	-1.4929	0.0439	-0.0048
$n=500$	β_{11}	1.2980	0.0440	-0.0016	$n=500$	β_{11}	1.3024	0.0442	0.0019
	β_{21}	1.4953	0.0450	-0.0031		β_{21}	1.5005	0.0449	0.0004
	θ	0.0519	0.0500	-0.4812		θ	0.2712	0.1730	1.7119
$m=30$	β_0	0.0037	0.2508	-0.9633	$m=30$	β_0	0.2748	0.2638	1.7475
$n=10$	β_{11}	-1.5004	0.2664	0.0002	$n=10$	β_{11}	-1.5353	0.2779	0.0235
	β_{21}	1.2932	0.2649	-0.0052		β_{21}	1.3236	0.2770	0.0182
	θ	1.4969	0.0974	-0.0021		θ	1.5328	0.2743	0.0219
$m=30$	β_0	0.0026	0.1460	-0.9737	$m=30$	β_0	0.2537	0.1477	1.5371
$n=30$	β_{11}	-1.4780	0.1544	-0.0147	$n=30$	β_{11}	-1.5052	0.1565	0.0035
	β_{21}	1.2760	0.1559	-0.0185		β_{21}	1.2994	0.1574	-0.0005
	θ	1.4809	0.0984	-0.0127		θ	1.5089	0.2088	0.0059
$m=30$	β_0	0.0016	0.1095	-0.9838	$m=30$	β_0	0.2699	0.1083	1.6987
$n=50$	β_{11}	-1.4717	0.1164	-0.0189	$n=50$	β_{11}	-1.4983	0.1165	-0.0011
	β_{21}	1.2764	0.1173	-0.0181		β_{21}	1.2996	0.1178	-0.0003
	θ	1.4775	0.0986	-0.0150		θ	1.5049	0.1987	0.0033
$m=30$	β_0	0.0014	0.0839	-0.9865	$m=30$	β_0	0.2811	0.0812	1.8110
$n=100$	β_{11}	-1.4728	0.0882	-0.0181	$n=100$	β_{11}	-1.4999	0.0874	-0.0001
	β_{21}	1.2813	0.0866	-0.0144		β_{21}	1.3050	0.0846	0.0038
	θ	1.4736	0.0983	-0.0176		θ	1.5013	0.1994	0.0009
$m=30$	β_0	0.0019	0.0349	-0.9815	$m=30$	β_0	0.2936	0.0348	1.9358
$n=500$	β_{11}	-1.4956	0.0381	-0.0030	$n=500$	β_{11}	-1.4966	0.0381	-0.0023
	β_{21}	1.2966	0.0379	-0.0027		β_{21}	1.3007	0.0377	0.0005
	θ	1.4945	0.0351	-0.0037		θ	1.4993	0.1989	-0.0004

Figure 1 displays the MSEs of the average estimated random effects estimates ($\widehat{\boldsymbol{u}}$) for each combination of (q, n_q). For easiness, we considered an equal number of sample sizes for each small area. Based on the simulation results, it shows that the fixed effects estimates from the proposed CPH with bias correction method provides equal or slightly better results compared to GLMM except for for the case $(5,10)$. The RMSE and Bias values are approximately similar or better in the proposed CPH approach in most cases. Most importantly, it is clear that the MHLE of variance parameter is more consistent and better regardless of the sample size. The simulation results are comparable with key findings in the literature which is proved that estimations based on h-likelihood provide equal or better results compared to other modeling approaches. When m increases, by asymptotic properties of estimators, it will not increase the complexity of parameter estimation process.

Figure 1: Mean Squares Error (MSE) of random effects estimates with samples of sizes 10, 30, 50, 100, and 500 with 5, 10, 20, and 30 small areas.



7.2 Simulation – Joint Modeling

Chapter 8. Real Data Analysis

8.1 CPH with Bias Correction Approach on Tobacco Smoking Data

We illustrate the proposed approach using a real data set of tobacco smoking combined 2015 and 2017 from Behavioral Risk Factor Surveillance System (BRFSS) which includes four variables of interest: ever use of E-cigarettes (EE), current use of E-cigarettes (CE), ever smoke (ES), and current smoke (CS) with sample size of 29404, 28162, 25711, and 29396 respectively. We apply the proposed h-likelihood method on BRFSS data set to incorporate individual-level tobacco use behaviors and area-level (county and state) ecological characteristics to electronic use prevalence among youth at the community level (i.e., county). The auxiliary information is considered for $\text{age}_i, i = 1, 2, 3$ ($<= 14, 15-17, >= 18$), race (4 groups: white, African American, Hispanic, and Others), sex (2 groups: male and female), year (2 groups: 2015, 2017) and poverty values. The data is available for 95 states. The poverty information is extracted from the US census between 2015 and 2017. Table 2 presents the summary statistics for each response variable.

The model is applied to all four variables of interest $(\boldsymbol{y}_{EE}, \boldsymbol{y}_{CE}, \boldsymbol{y}_{ES}, \boldsymbol{y}_{CS})$ separately. The MHLEs of fixed effects, $\widehat{\boldsymbol{\beta}}_{EE}, \widehat{\boldsymbol{\beta}}_{CE}, \widehat{\boldsymbol{\beta}}_{ES}$, and $\widehat{\boldsymbol{\beta}}_{CS}$ are obtained using the proposed CPH with bias correction approach and the benchmark analysis based on GLMM. The tables 7.1, 7.2, 7.3, and 7.4 displays the fixed effects estimates obtained from each method.

Table 8.1: Model estimates for current use of E-Cigarettes based on the proposed method and GLMM

E-Cigarrete Current									
		CPH w/t Bias Correction				GLMM			
		Estimate	**SE**	**Z Value**	**P(>|Z|)**	**Estimate**	**SE**	**Z Value**	**P(>|Z|)**
Intercept		-1.625	0.132	-12.290	0.000	-1.877	0.181	-10.356	0.000
Age	<= 14 yrs	0.000				0.000			
	15-17 yrs	0.305	0.053	5.751	0.000	0.324	0.086	3.786	0.000

	>= 18 yrs	0.688	0.065	10.633	0.000	0.691	0.086	8.006	0.000
Race	White	0.212	0.056	3.788	0.000	0.245	0.084	2.927	0.003
	African American	-0.243	0.073	-3.345	0.001	-0.176	0.105	-1.674	0.094
	Hispanic	0.142	0.059	2.409	0.016	0.193	0.092	2.102	0.036
	Others	0.000				0.000			
Sex	Male	0.000				0.000			
	Female	0.291	0.031	9.270	0.000	0.250	0.045	5.499	0.000
Year	2015	0.000				0.000			
	2017	-0.946	0.058	-16.255	0.000	-0.794	0.077	-10.365	0.000
Poverty Rate (%)		-0.564	0.683	-0.825	0.409	-0.090	0.897	-0.100	0.920

Table 8.2: Model estimates for ever use of E-Cigarettes based on the proposed method and GLMM

E-Cigarrete Ever									
		CPH w/t Bias Correction				**GLMM**			
		Estimate	**SE**	**Z Value**	**P(>|Z|)**	**Estimate**	**SE**	**Z Value**	**P(>|Z|)**
Intercept		-0.905	0.101	-8.988	0.000	-0.990	0.144	-6.880	0.000
Age	<= 14 yrs	0.000				0.000			
	15-17 yrs	0.447	0.039	11.413	0.000	0.407	0.060	6.736	0.000
	>= 18 yrs	0.767	0.050	15.307	0.000	0.771	0.062	12.421	0.000
Race	White	0.095	0.042	2.245	0.025	0.065	0.062	1.038	0.299
	African American	-0.073	0.051	-1.418	0.156	-0.018	0.077	-0.232	0.817
	Hispanic	0.311	0.044	7.040	0.000	0.349	0.068	5.098	0.000
	Others	0.000				0.000			
Sex	Male	0.000				0.000			
	Female	0.137	0.024	5.644	0.000	0.127	0.035	3.612	0.000
Year	2015	0.000				0.000			
	2017	-0.172	0.043	-4.010	0.000	-0.106	0.061	-1.751	0.080
Poverty Rate (%)		0.575	0.522	1.101	0.271	0.653	0.739	0.883	0.377

Table 8.3: Model estimates for current smoke based on the proposed method and GLMM

Current Smoke									
		CPH w/t Bias Correction				**GLMM**			
		Estimate	**SE**	**Z Value**	**P(>|Z|)**	**Estimate**	**SE**	**Z Value**	**P(>|Z|)**
Intercept		-3.337	0.171	-19.500	0.000	-3.412	0.240	-14.206	0.000
Age	<= 14 yrs	0.000				0.000			
	15-17 yrs	0.429	0.075	5.732	0.000	0.383	0.120	3.189	0.001
	>= 18 yrs	1.099	0.086	12.776	0.000	0.908	0.119	7.605	0.000

Race	White	0.290	0.074	3.907	0.000	0.187	0.107	1.752	0.080
	African American	-0.615	0.102	-6.028	0.000	-0.591	0.146	-4.062	0.000
	Hispanic	0.087	0.080	1.087	0.277	-0.015	0.121	-0.123	0.902
	Others	0.000				0.000			
Sex	Male	0.000				0.000			
	Female	0.235	0.041	5.721	0.000	0.174	0.058	2.995	0.003
Year	2015	0.000				0.000			
	2017	-0.261	0.073	-3.561	0.000	-0.198	0.099	-1.997	0.046
Poverty Rate (%)		2.655	0.843	3.151	0.002	3.297	1.151	2.864	0.004

Table 8.4: Model estimates for ever smoke based on the proposed method and GLMM

Ever Smoke									
		CPH w/t Bias Correction				**GLMM**			
		Estimate	**SE**	**Z Value**	**P(>|Z|)**	**Estimate**	**SE**	**Z Value**	**P(>|Z|)**
Intercept		-1.939	0.138	-14.005	0.000	-1.868	0.188	-9.945	0.000
Age	<= 14 yrs	0.000				0.000			
	15-17 yrs	0.477	0.052	9.190	0.000	0.337	0.071	4.732	0.000
	>= 18 yrs	1.002	0.060	16.576	0.000	0.857	0.072	11.920	0.000
Race	White	0.177	0.050	3.573	0.000	0.008	0.069	0.120	0.905
	African American	-0.302	0.062	-4.835	0.000	-0.278	0.086	-3.242	0.001
	Hispanic	0.217	0.051	4.246	0.000	0.165	0.075	2.197	0.028
	Others	0.000				0.000			
Sex	Male	0.000				0.000			
	Female	0.136	0.028	4.813	0.000	0.100	0.038	2.606	0.009
Year	2015	0.000				0.000			
	2017	-0.317	0.053	-6.022	0.000	-0.235	0.075	-3.129	0.002
Poverty Rate (%)		3.688	0.737	5.003	0.000	3.811	1.004	3.794	0.000

The random effects for missing counties are obtained using the nearest neighboring approach assuming that the effects of neighboring areas are correlated, with correlation decaying to zero as distance increases. The estimated random effect for missing county $c_i(\tilde{\mu}_{c_i})$

$$\tilde{\mu}_{c_i} = \hat{\mu}_{c_j}, s.t. min\,\text{dist}(c_i, c_k)\,,\ k = 1, \dots, m-1,$$

where c_j is the closest to county c_i. The unit-level prevalence for each combination $(3 \times 2 \times 2 \times 4)$ of groups are estimated using

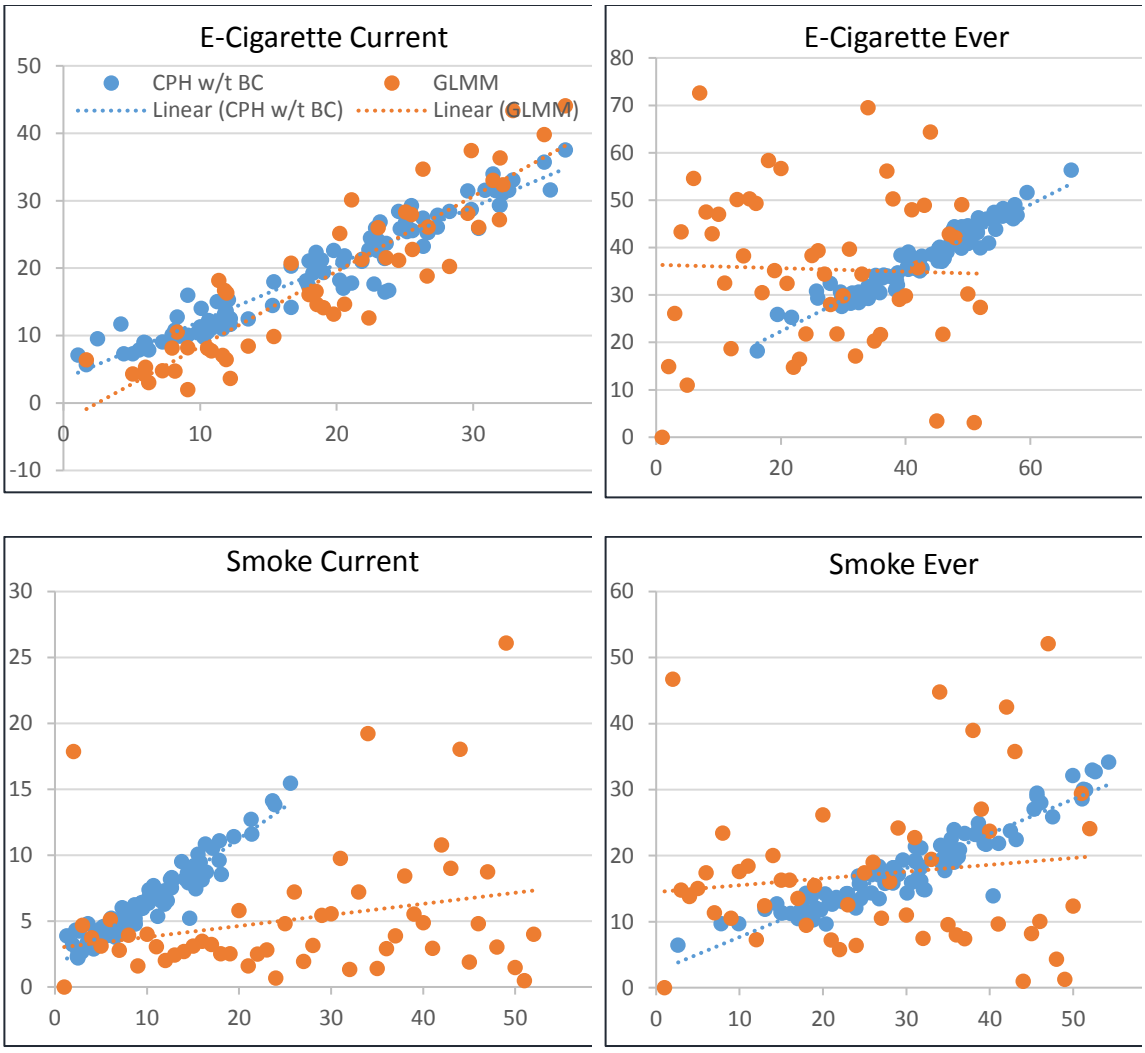
$$\widetilde{\mathrm{P}}_{ijkc}\left(y_{ijkc} = 1 \middle| u_c\right) = \frac{\exp\left(\hat{\alpha}_i + \hat{\beta}_j + \hat{\gamma}_k + x_c'\hat{\eta} + \hat{u}_c\right)}{1 + \exp\left(\hat{\alpha}_i + \hat{\beta}_j + \hat{\gamma}_k + x_c'\hat{\eta} + \hat{u}_c\right)},$$

where $\hat{\alpha}_i(i = 1,2,3), \hat{\beta}_j(1,2), \hat{\gamma}_k(k = 1,2,3,4)$, and $\hat{\eta}$ are coefficient estimates for age, gender, race, and poverty rate respectively. The unit-level estimates were used to obtain the county-level estimates using the U.S. Census population as in equation (8).

$$\widetilde{\mathrm{P}}(y_c = 1|u) = \frac{\sum_i \sum_j \sum_k \tilde{P}_{ijkc} \times \mathrm{Pop}_{ijkc}}{\mathrm{Pop}_c}, \tag{8}$$

where $\mathrm{Pop}_c = \sum_i \sum_j \sum_k \mathrm{Pop}_{ijkc}$ is the total population for county c. The model predicted prevalence and observed prevalence were compared using Pearson's and Spearmen's correlation coefficients. The predicted proportions for "ever use of E-cigarettes", "current use of E-cigarettes", "ever-smoke", and "current-smoke" are obtained using the proposed method through MHLEs. Figure 3**Error! Reference source not found.** shows the estimated county prevalence for each variable of interest.

Figure 2. Observed and estimated prevalence ever-use and current-use of E-Cigarettes, and current smoke and ever smoke in the United States based on 2015-2017 YRBSS data



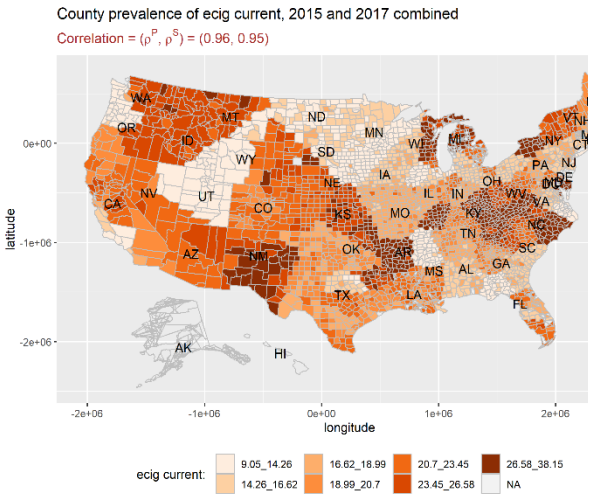
Both Spearman's and Pearson's correlation coefficients ranged from 0.93 to 0.96 indicate that the observed and estimated county prevalences are reliable and better. The results show that the prevalence of ever use of E-cigarettes is higher than the prevalence of current use of E-cigarettes, ever smoke, and current smoke. The results indicate that the current use of E-cigarettes among youth is higher than current smoking in most US counties. It is also clear that the counties with the lower and higher prevalence of the current use of E-cigarettes are more likely to have lower and higher prevalence for ever use of E-

cigarettes. Similarly, the areas with lower and higher prevalence for current smoke are more likely to have lower and higher prevalence for ever smoke.

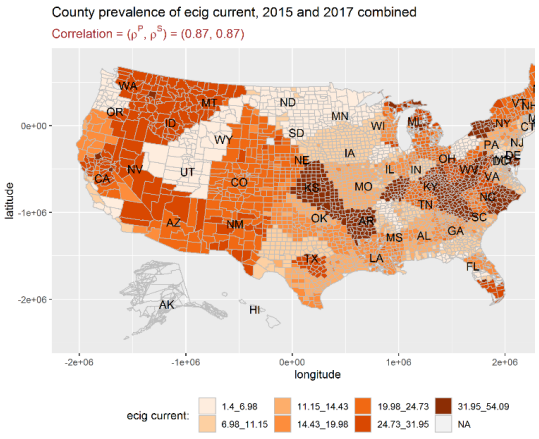
Table 5 and Table 6 present the 10 US states with the lowest and highest estimated prevalence for each response variable. It shows that West Virginia and Kentucky states have the highest prevalence of current smoking (11.89%, and 10.28%), North Carolina and Kentucky have the highest prevalence of current use of E-cigarettes (29.22%, and 27.38%), respectively. New Mexico has the highest prevalence for both ever smoke and ever use of E-cigarettes among all the states in the U.S. The estimated prevalence decreases from the year 2015 to 2017 for all four variables of interest. Table 7 presents the estimated prevalence for each U.S. state using the proposed CPH with bias correction and generalized linear mixed model for current use, ever use of E-Cigarettes, current smoke, and ever smoke.

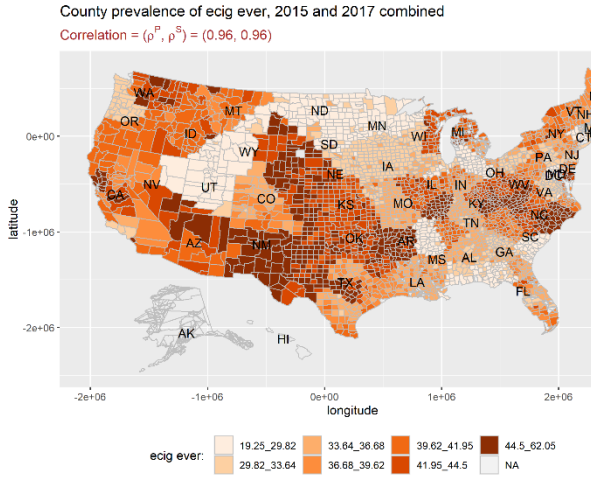
Figure 3. County prevalence of ever-use and current-use of E-Cigarettes in the United States based on 2015-2017 YRBSS data

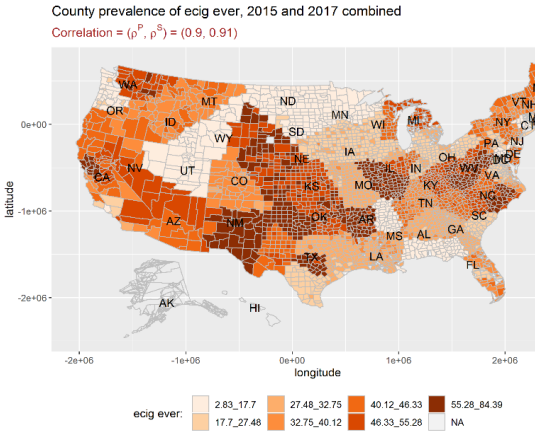
(a) CPH

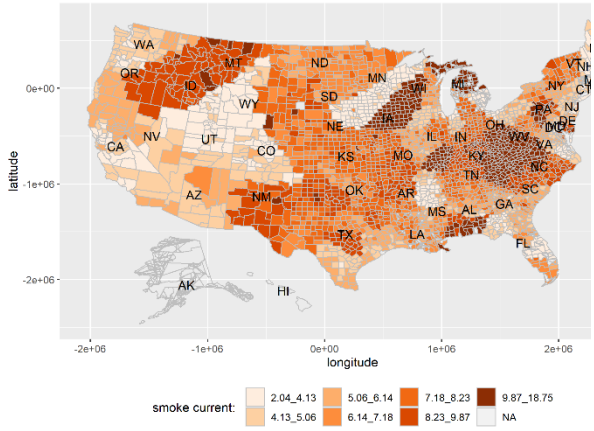


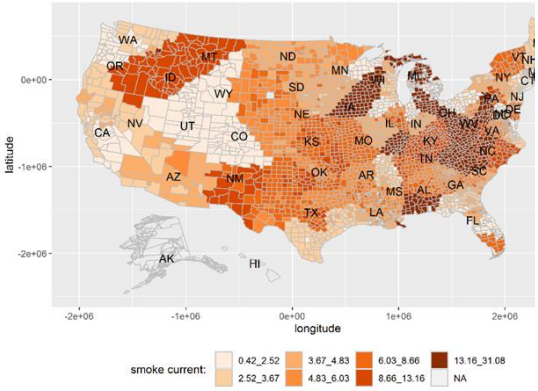
(b) GLMM

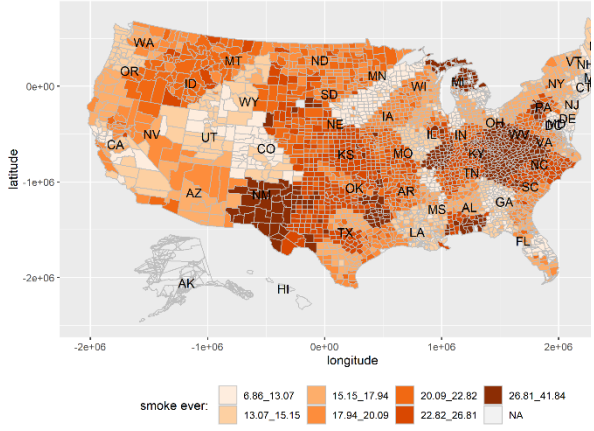


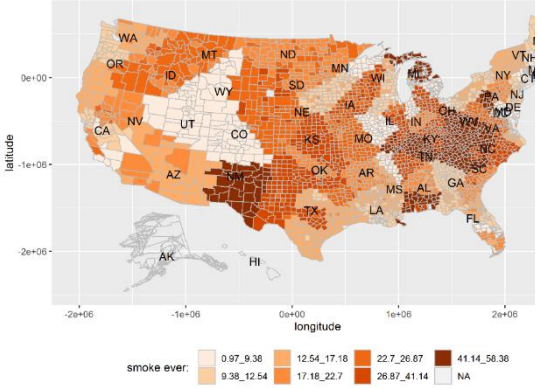

County prevalence of ecig ever, 2015 and 2017 combined
Correlation = (ρ^P, ρ^S) = (0.96, 0.96)
latitude
longitude
ecig ever:
19.25_29.82
29.82_33.64
33.64_36.68
36.68_39.62
39.62_41.95
41.95_44.5
44.5_62.05
NA


County prevalence of ecig ever, 2015 and 2017 combined
Correlation = (ρ^P, ρ^S) = (0.9, 0.91)
latitude
longitude
ecig ever:
2.83_17.7
17.7_27.48
27.48_32.75
32.75_40.12
40.12_46.33
46.33_55.28
55.28_84.39
NA


County prevalence of smoke current, 2015 and 2017 combined
Correlation = (ρ^P, ρ^S) = (0.96, 0.95)
latitude
longitude
smoke current:
2.04_4.13
4.13_5.06
5.06_6.14
6.14_7.18
7.18_8.23
8.23_9.87
9.87_18.75
NA


County prevalence of smoke current, 2015 and 2017 combined
Correlation = (ρ^P, ρ^S) = (0.9, 0.92)
latitude
longitude
smoke current:
0.42_2.52
2.52_3.67
3.67_4.83
4.83_6.03
6.03_8.66
8.66_13.16
13.16_31.08
NA


County prevalence of smoke ever, 2015 and 2017 combined
Correlation = (ρ^P, ρ^S) = (0.94, 0.93)
latitude
longitude
smoke ever:
6.86_13.07
13.07_15.15
15.15_17.94
17.94_20.09
20.09_22.82
22.82_26.81
26.81_41.84
NA


County prevalence of smoke ever, 2015 and 2017 combined
Correlation = (ρ^P, ρ^S) = (0.87, 0.88)
latitude
longitude
smoke ever:
0.97_9.38
9.38_12.54
12.54_17.18
17.18_22.7
22.7_26.87
26.87_41.14
41.14_58.38
NA

8.2 Joint Modeling of Multiple Outcomes based on Tobacco Smoking Data

We evaluate the proposed joint modeling approach using the same tobacco smoking data set described above considering current use of E-Cigarretes and ever use of E-Cigarretes. The association between current use and ever use of E-Cigarretes had been examined through the shared parameter.

Chapter 9. Discussion

The proposed CPH approach is computationally efficient and provides accurate parameter estimates for fixed effects, random effects, and variance components using h-likelihood through numerical approximation technique. For simplicity, we only considered $u_i \sim \mathcal{N}(0, \sigma^2), i = 1, \dots, m$, but random effects may come from any conjugate of exponential family distributions in HGLMs and can directly estimate model parameters using the proposed CPH method. The h-likelhood approach avoids computationally expensive integration by taking partial derivatives of logarithm of joint density function, namely h-likelihood which simplifies the differentiation. Moreover, the proposed h-likelihood approach is a computationally efficient method that provides reliably accurate results through a single algorithm.

The proposed h-likelihood method with bias correction of estimates provides consistently better estimates compared to the h-likelihood method without bias correction. The simulation results demonstrate that this method performs well providing reliable estimates even for the areas with small sample sizes, which are given in Table 1. Even though the MHLEs of variance parameter σ^2 provide better estimates with a large number of small areas, it also provides reliable estimates

with small sample sizes and with a small number of small areas. Similarly, the MHLEs of fixed effects are reasonably accurate in any scenario, but they are more accurate with large samples compared to small sample sizes and a small number of small areas. Overall, the proposed CPH model results are consistent, less complex parameter estimation, and slightly better in some cases compared with the related historical work done by the other researchers (Breslow & Clayton, 1993; Shun & McCullagh, 1995). As we expected, the mean squared error of fixed effects estimates gets smaller when increasing the sample size as well as the number of small areas. Overall, the h-likelihood produces reasonable MHLEs. In some situations, there might be convergence issues with very small sample sizes with a small number of areas, especially with binary response data.

The mean squared errors obtained from MC simulation for the average random effects are displayed in Figure 1 for each combination of the number of small areas and sample sizes. Overall, it shows that the MSE decreases when the sample size increases. In this article, we assumed that the estimates of $\boldsymbol{\beta}$ and $\boldsymbol{u}$ are independent of variance components $\theta(=\sigma^2)$, and $\hat{\sigma}^2$ is obtained using partial derivatives. Practically, it might be the case where $\boldsymbol{\beta}$ and $\boldsymbol{u}$ are functions of σ^2, so estimation of σ^2 expected to be more accurate based on the total derivative of h_A with respect to some σ^2 to avoid the indirect independencies between the estimators. The future work will be based on total derivative to estimate σ^2.

The proposed method has opportunities to extend in other applications in SAE. With this, we will also consider extending this methodology for high dimensional case in SAE. In this article, we consider two models for multiple outcomes (current use, ever use of E-Cigarettes, current smoke, and ever smoke), but we plan to extend the proposed methodology to joint modeling to account for the association between multiple outcomes that occur in the same small area.

Appendix

Appendix 1. Results for ever use of E-cigarettes based on NYTS data

Table 1. Predicted versus true prevalence for ever-use of E-cigarettes using GLMM model

County	Predicted Prevalence	Observed Prevalence	County	Predicted Prevalence	Observed Prevalence
1013	19.78	22.67	**31053**	19.19	17.82
1081	26.83	26.06	**32003**	32.45	26.25
1097	21.41	20.58	**33013**	11.88	13.51
4013	23.05	21.75	**34003**	13.00	11.24
4021	19.69	18.99	**34007**	16.95	16.18
4027	18.82	18.52	**34017**	7.70	7.98
5031	34.51	31.58	**34023**	22.20	15.38
5037	21.81	23.91	**34029**	27.36	30.06
5131	36.08	41.42	**34031**	38.93	29.90
6019	14.24	16.37	**34039**	13.73	26.67
6037	27.49	26.34	**36047**	7.78	9.71
6043	8.76	9.20	**36067**	17.43	12.05
6059	13.11	13.04	**36071**	29.26	35.54
6065	23.80	22.93	**36081**	20.55	10.98
6071	29.05	30.38	**36085**	20.28	22.63
6073	26.04	26.46	**36087**	11.8	14.40
6083	33.74	35.64	**36103**	38.00	33.33
6085	17.21	17.13	**36119**	30.43	40.77
9011	8.59	4.26	**37081**	24.05	19.76
12009	36.96	26.5	**37119**	19.46	19.63
12011	20.65	25.99	**37191**	26.52	29.61
12019	21.65	20.70	**38017**	20.70	14.48
12021	28.81	20.61	**39017**	35.55	20.41
12057	21.85	23.46	**39035**	13.79	8.62
12071	33.19	33.55	**39081**	40.29	40.00
12081	29.29	31.69	**39151**	22.67	23.08
12086	29.42	18.23	**39171**	17.36	10.53
12095	14.90	14.44	**40027**	34.33	30.93
12097	7.33	7.72	**40039**	36.66	42.19
12099	20.08	19.09	**40109**	63.46	72.97
12111	12.73	8.40	**40137**	28.52	22.86
12117	14.68	9.15	**41029**	26.45	36.36
13063	16.17	16.26	**41035**	31.99	12.08
13095	16.87	14.29	**42003**	17.59	23.34
13175	13.05	16.25	**42041**	6.51	2.60
13245	30.87	32.63	**42073**	34.78	45.83
15001	53.40	43.66	**42085**	30.05	32.55

16027	26.25	26.29	**42091**	10.99	20.00
17031	24.14	16.93	**42101**	17.77	17.39
17043	19.95	16.03	**45071**	22.34	22.04
17091	26.86	25.74	**46005**	22.03	20.72
17097	29.10	30.70	**46011**	15.27	13.77
17113	8.42	14.29	**46103**	26.55	19.21
17119	19.79	10.14	**47053**	31.73	33.56
17197	22.87	15.60	**47119**	23.12	27.06
17201	28.77	25.35	**47157**	23.71	30.00
18097	24.94	19.60	**47165**	26.47	22.92
19065	10.91	13.07	**48027**	17.80	15.18
19113	20.92	30.43	**48029**	24.59	26.46
21111	30.60	50.91	**48113**	26.44	16.48
22063	19.17	21.13	**48141**	42.87	42.20
22071	12.48	6.31	**48201**	19.60	24.59
24003	28.55	28.87	**48215**	23.07	23.13
24005	8.92	4.52	**48303**	27.61	39.58
24031	16.83	12.85	**48375**	30.97	42.19
25005	31.81	41.6	**48439**	32.14	28.65
25009	13.38	15.79	**49049**	14.72	15.70
25017	17.97	19.32	**51041**	26.85	27.81
25027	26.19	26.06	**51051**	14.81	15.79
26045	6.52	5.75	**51095**	20.97	15.24
26049	23.53	21.61	**51740**	18.64	11.54
26099	17.22	15.58	**53033**	18.92	28.57
26125	26.75	35.45	**53077**	21.92	17.34
26133	9.73	12.77	**54003**	20.61	24.59
26163	17.99	17.40	**54049**	23.63	24.21
27003	23.89	26.75	**55025**	17.78	20.78
28085	32.65	32.51	**55079**	13.95	17.27
28107	21.13	19.72	**55121**	8.76	4.62
29037	19.70	19.81	**55127**	7.98	4.29
29095	13.50	8.67	**55133**	24.18	31.15
30063	21.82	21.39	**55139**	24.29	12.16

Appendix 2. Results for current use of E-cigarettes based on NYTS data

Table 2. Predicted Vs. true prevalence for current use of E-cigarettes using GLMM model

County	Predicted Prevalence	Observed Prevalence	County	Predicted Prevalence	Observed Prevalence
1013	7.03	8.64	31053	9.86	8.91
1081	13.06	12.59	32003	14.56	12.14
1097	14.42	14.15	33013	3.42	4.30
4013	9.17	8.86	34003	6.97	5.49
4021	8.38	7.80	34007	4.92	5.26
4027	6.23	6.42	34017	1.97	1.67
5031	20.14	16.37	34023	5.08	5.64
5037	6.01	6.44	34029	13.68	15.61
5131	18.83	21.36	34031	18.42	13.27
6019	7.01	7.44	34039	5.25	9.60
6037	9.99	9.69	36047	0.65	1.47
6043	4.35	5.20	36067	7.09	3.12
6059	4.80	4.40	36071	8.13	9.52
6065	8.11	9.05	36081	5.24	2.42
6071	11.24	10.91	36085	10.97	12.45
6073	12.54	13.42	36087	3.93	5.33
6083	12.63	14.29	36103	14.76	11.98
6085	8.96	8.29	36119	8.65	10.77
9011	4.95	2.15	37081	10.71	8.10
12009	15.89	11.97	37119	8.84	8.79
12011	8.57	10.48	37191	15.45	17.27
12019	10.32	10.75	38017	8.53	5.80
12021	10.60	7.63	39017	16.94	8.16
12057	8.04	9.66	39035	3.36	2.54
12071	17.19	17.36	39081	19.96	19.48
12081	16.26	16.80	39151	12.24	8.24
12086	17.10	9.84	39171	3.33	1.75
12095	6.49	5.88	40027	20.08	17.72
12097	2.27	2.46	40039	13.67	14.12
12099	7.58	7.44	40109	44.31	57.89
12111	4.65	2.56	40137	9.61	6.38
12117	8.60	4.85	41029	8.56	12.87
13063	7.53	8.31	41035	1.14	4.00
13095	9.95	8.24	42003	7.96	12.57
13175	2.85	3.26	42073	24.56	36.11
13245	13.62	15.38	42085	13.36	13.21
15001	32.09	26.76	42091	4.44	7.25
16027	13.65	13.82	42101	5.52	5.00
17031	11.76	7.55	45071	5.64	5.57

17043	12.18	9.51	46005	6.93	7.14
17091	8.86	9.17	46011	4.94	4.96
17097	18.16	18.67	46103	11.48	7.87
17113	3.52	5.56	47053	10.10	12.03
17119	11.81	5.46	47119	12.64	15.13
17197	13.57	8.33	47157	8.14	12.2
17201	16.61	15.46	47165	12.49	8.85
18097	5.77	6.09	48027	6.79	6.21
19065	2.58	3.98	48029	12.80	13.5
19113	7.00	11.07	48113	19.01	9.78
21111	7.32	17.73	48141	22.03	23.17
22063	5.73	6.97	48201	11.40	14.57
22071	0.03	0.89	48215	8.04	7.46
24003	11.19	11.83	48303	16.97	27.08
24005	6.05	3.23	48375	8.62	12.31
24031	7.85	5.58	48439	9.68	10.22
25005	18.92	23.20	49049	6.53	7.18
25009	4.12	5.24	51041	13.96	15.13
25017	4.57	4.89	51051	1.60	1.59
25027	15.29	15.41	51095	12.02	8.92
26045	4.47	4.60	51740	10.40	4.67
26049	10.56	9.57	53033	9.41	16.07
26099	6.23	5.10	53077	6.64	5.04
26125	11.07	14.93	54003	7.57	10.61
26133	4.00	5.32	54049	13.78	13.20
26163	5.49	5.21	55025	9.69	10.68
27003	10.51	12.08	55079	6.33	7.77
28085	18.66	18.88	55121	2.19	1.52
28107	5.54	5.19	55127	7.19	2.94
29037	7.94	7.50	55133	7.25	11.48
29095	3.35	2.63	55139	8.95	4.76
30063	14.64	13.87			

Appendix 3. Simulation Results for Mixed Logit Model Based On CPH with Bias Correction Approach

Appendix 4. R program for CPH with Bias Correction Approach in SAE

The proposed CPH with bias correction approach is illustrated through a simulation study and a real data set for the mixed logit model using R programming language. The function “cph.fit”

function esitmates the fixed effects, random effects, and variance parameter based on the proposed method.

Usage

The matrix form of each distribution function: $P(\boldsymbol{y}|\boldsymbol{u}) = (1 + \exp(-\boldsymbol{X\beta} - \boldsymbol{Zu}))^{-1}$ and $f(\boldsymbol{u}) = (2\pi)^{-m/2}(det(\boldsymbol{G}))^{-1/2} \exp\left(-\frac{1}{2}\boldsymbol{u}^T\boldsymbol{G^{-1}u}\right)$.

From (4.3), the h-likelihood function can be expressed as

$$h = \boldsymbol{y}^T(\boldsymbol{X\beta} + \boldsymbol{Zu}) - \mathbf{1}^T \log(1 + \exp(\boldsymbol{X\beta} + \boldsymbol{Zu})) - \frac{m}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\boldsymbol{u}^T\boldsymbol{u} + c,$$

$$h = \boldsymbol{y}^T(\boldsymbol{X\beta} + \boldsymbol{Zu}) - \mathbf{1}^T \log(1 + \exp(\boldsymbol{X\beta} + \boldsymbol{Zu})) - \frac{1}{2}\boldsymbol{u}^T\boldsymbol{G^{-1}}\boldsymbol{u} - \frac{1}{2}\log(\det(\boldsymbol{G})) + c,$$

where $\mathbf{1}^T$ is a unit vector with dimension $(1 \times N)$. Take partial derivative with respect to $\boldsymbol{\beta}$ and $\boldsymbol{u}$

$$\frac{\partial h}{\partial \boldsymbol{\beta}} = [\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{X}^T(1 + \exp(\boldsymbol{X\beta} + \boldsymbol{Zu}))^{-1}\exp(\boldsymbol{X\beta} + \boldsymbol{Zu})]_{p\times 1}$$

$$= \boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{X}^T\boldsymbol{\pi} = \boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{\pi}),$$

$$\frac{\partial h}{\partial \boldsymbol{u}} = [\boldsymbol{Z}^T\boldsymbol{y} - \boldsymbol{Z}^T(1 + \exp(\boldsymbol{X\beta} + \boldsymbol{Zu})^{-1}\exp(\boldsymbol{X\beta} + \boldsymbol{Zu}) - \boldsymbol{G}^{-1}\boldsymbol{u}]_{m\times 1}$$

$$= \boldsymbol{Z}^T\boldsymbol{y} - \boldsymbol{Z}^T\boldsymbol{\pi} - \boldsymbol{G}^{-1}\boldsymbol{u} = \boldsymbol{Z}^T(\boldsymbol{y} - \boldsymbol{\pi}) - \boldsymbol{G}^{-1}\boldsymbol{u},$$

where $\boldsymbol{\pi} = 1/(1 + \exp(-\boldsymbol{X\beta} - \boldsymbol{Zu}))$.

The components of the asymptotic variance-covariance matrix of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{u}}$, $\boldsymbol{\mathcal{J}}$ also known as the observed information matrix that is calculated by

$$\boldsymbol{\mathcal{J}} = \begin{bmatrix} -\frac{\partial^2 h}{\partial \boldsymbol{\beta}^2} & -\frac{\partial^2 h}{\partial \boldsymbol{\beta}\partial \boldsymbol{u}} \\ -\frac{\partial^2 h}{\partial \boldsymbol{u}\partial \boldsymbol{\beta}} & -\frac{\partial^2 h}{\partial \boldsymbol{u}^2} \end{bmatrix},$$

where

$$-\frac{\partial^2 h}{\partial \boldsymbol{\beta}^2} = \frac{\partial}{\partial \boldsymbol{\beta}}(\boldsymbol{X}^T(1 + \exp(-\boldsymbol{X\beta} - \boldsymbol{Zu}))^{-1}\exp(-\boldsymbol{X\beta} - \boldsymbol{Zu})\boldsymbol{X})$$

$$= \boldsymbol{X}^T\frac{1}{(1 + \exp(-\boldsymbol{X\beta} - \boldsymbol{Zu}))}\left(\frac{1}{1 + \exp(\boldsymbol{X\beta} + \boldsymbol{Zu})}\right)\boldsymbol{X}$$

$$= \boldsymbol{X}^T\boldsymbol{\pi}(\mathbf{1} - \boldsymbol{\pi})\boldsymbol{X}$$

$$= \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X},$$

$$-\frac{\partial^2 h}{\partial \boldsymbol{\beta}\partial \boldsymbol{u}} = \boldsymbol{X}^T\frac{\partial}{\partial \boldsymbol{u}}((1 + \exp(-\boldsymbol{X\beta} - \boldsymbol{Zu}))^{-1}\exp(-\boldsymbol{X\beta} - \boldsymbol{Zu}))$$

$$= \boldsymbol{X}^T(1 + \exp(-\boldsymbol{X\beta} - \boldsymbol{Zu}))^{-2}\boldsymbol{Z}$$

$$= \boldsymbol{X}^T\boldsymbol{\pi}(\mathbf{1} - \boldsymbol{\pi})\boldsymbol{Z}$$

$$= \boldsymbol{X}^T\boldsymbol{WZ},$$

$$-\frac{\partial^2 h}{\partial \boldsymbol{u}\partial \boldsymbol{\beta}} = -\frac{\partial}{\partial \boldsymbol{\beta}}(\boldsymbol{Z}^T\boldsymbol{y} - \boldsymbol{Z}^T(1 + \exp(-\boldsymbol{X\beta} - \boldsymbol{Zu}))^{-1} - \boldsymbol{G}^{-1}\boldsymbol{u})$$
$$= \boldsymbol{Z}^T(1 + \exp(-\boldsymbol{X\beta} - \boldsymbol{Zu}))^{-2}\boldsymbol{X}$$
$$= \boldsymbol{Z}^T\boldsymbol{\pi}(\mathbf{1} - \boldsymbol{\pi})\boldsymbol{X}$$
$$= \boldsymbol{Z}^T\boldsymbol{WX},$$

$$-\frac{\partial^2 h}{\partial \boldsymbol{u}^2} = -\frac{\partial}{\partial \boldsymbol{u}}(\boldsymbol{Z}^T\boldsymbol{y} - \boldsymbol{Z}^T(1 + \exp(-\boldsymbol{X\beta} - \boldsymbol{Zu}))^{-1} - \boldsymbol{G}^{-1}\boldsymbol{u})$$

$$= \boldsymbol{Z}^T(1 + \exp(-\boldsymbol{X\beta} - \boldsymbol{Zu}))^{-2}\boldsymbol{Z} + \boldsymbol{G}^{-1}$$

$$= \boldsymbol{Z}^T\boldsymbol{\pi}(\mathbf{1} - \boldsymbol{\pi})\boldsymbol{Z} + \boldsymbol{G}^{-1}$$
$$= \boldsymbol{Z}^T\boldsymbol{WZ} + \boldsymbol{G}^{-1},$$

where $\boldsymbol{W}$ is a $N \times N$ diagonal matrix with the diagonal elements of each block being $\boldsymbol{\pi}_i(\mathbf{1} - \boldsymbol{\pi}_i)$ for area i.

References

ACS. Bureau of the Census, American Community Survey (ACS). Retrieved from https://www.census.gov/data/developers/data-sets.html

Akaike, H. (1973). Information theory and the maximum likelihood principle in 2nd International Symposium on Information Theory (BN Petrov and F. Cs ä ki, eds.). *Akademiai Ki à do, Budapest*.

Arnab, R. (2017). *Survey sampling theory and applications*: Academic Press.

Barrington-Trimis, J. L., Urman, R., Berhane, K., Unger, J. B., Cruz, T. B., Pentz, M. A., . . . McConnell, R. (2016). E-Cigarettes and Future Cigarette Use. *Pediatrics*. doi:10.1542/peds.2016-0379

Battese, G. E., & Fuller, W. A. (1981). *Prediction of county crop areas using survey and satellite data.* Paper presented at the Proceedings of the section on survey research methods, American Statistical Association.

Berkowitz, Z., Zhang, X., Richards, T. B., Peipins, L., Henley, S. J., Holt, J. J. C. E., & Biomarkers, P. (2016). Multilevel small-area estimation of multiple cigarette smoking status categories using the 2012 behavioral risk factor surveillance system. *25*(10), 1402-1410.

Berkowitz, Z., Zhang, X., Richards, T. B., Sabatino, S. A., Peipins, L. A., Holt, J., & White, M. C. (2019). Multilevel Regression for Small-Area Estimation of Mammography Use in the United States, 2014. *Cancer Epidemiol Biomarkers Prev, 28*(1), 32-40. doi:10.1158/1055-9965.EPI-18-0367

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American statistical association, 88*(421), 9-25.

Cameron, J. M., Howell, D. N., White, J. R., Andrenyak, D. M., Layton, M. E., & Roll, J. M. (2014). Variable and potentially fatal amounts of nicotine in e-cigarette nicotine solutions. *Tob Control, 23*(1), 77-78. doi:10.1136/tobaccocontrol-2012-050604

Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: a modern perspective*: Chapman and Hall/CRC.

Casella, G., & Berger, R. L. (2002). *Statistical inference* (Vol. 2): Duxbury Pacific Grove, CA.

Centers for Disease Control and Prevention. (2013). QuickStats: Number of Deaths from 10 Leading Causes—National Vital Statistics System, United States, 2010. . *Morbidity and Mortality Weekly Report, http://www.cdc.gov/mmwr/preview/mmwrhtml/mm6325a3.htm?s_cid=mm6325a3_w, 62*(08), 155.

Cressie, N. (1991). *Small-area prediction of undercount using the general linear model.* Paper presented at the Proceedings of statistics symposium 90: measurement and improvement of data quality.

Dai, H. (2018). Single, Dual, and Poly Use of Flavored Tobacco Products among Youth in the United States: 2014. *Preventing Chronic Disease. In press.*

Dai, H. (2019). Changes in Flavored Tobacco Product Use among Current Youth Tobacco Users in the United States: 2014 – 2017. *JAMA Pediatr*.

Dai, H., Catley, D., Richter, K. P., Goggin, K., & Ellerbeck, E. F. (2018). Electronic Cigarettes and Future Marijuana Use: A Longitudinal Study. *Pediatrics*. doi:10.1542/peds.2017-3787

Dai, H., Catley, D., Richter, K. P., Goggin, K., & Ellerbeck, E. F. J. P. (2018). Electronic Cigarettes and Future Marijuana Use: A Longitudinal Study. *141*(5), e20173787.

Daniels, M. J., & Gatsonis, C. (1999). Hierarchical generalized linear models in the analysis of variations in health care utilization. *Journal of the American statistical association, 94*(445), 29-42.

Datta, G. S., & Ghosh, M. J. T. A. o. S. (1991). Bayesian prediction in linear models: Applications to small area estimation. 1748-1770.

Dempster, A., & Tomberlin, T. (1980). *The analysis of census undercount from a postenumeration survey.* Paper presented at the Proceedings of the Conference on Census Undercount.

Dick, P. J. S. M. (1995). Modelling net undercoverage in the 1991 Canadian Census. *21*(1), 45-54.

Dinakar, C., & O'Connor, G. T. (2016). The Health Effects of Electronic Cigarettes. *N Engl J Med, 375*(14), 1372-1381. doi:10.1056/NEJMra1502466

Dwyer-Lindgren, L., Mokdad, A. H., Srebotnjak, T., Flaxman, A. D., Hansen, G. M., & Murray, C. J. (2014). Cigarette smoking prevalence in US counties: 1996-2012. *Population health metrics, 12*(1), 5.

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

Fay III, R. E., & Herriot, R. A. J. J. o. t. A. S. A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *74*(366a), 269-277.

Fuller, W. A., & Battese, G. E. (1973). Transformations for estimation of linear models with nested-error structure. *Journal of the American statistical association, 68*(343), 626-632.

Ghosh, M., Natarajan, K., Stroud, T., & Carlin, B. P. J. J. o. t. A. S. A. (1998). Generalized linear models for small-area estimation. *93*(441), 273-282.

Ghosh, M., & Rao, J. J. S. s. (1994). Small area estimation: an appraisal. *9*(1), 55-76.

Goniewicz, M. L., Hajek, P., & McRobbie, H. (2014). Nicotine content of electronic cigarettes, its release in vapour and its consistency across batches: regulatory implications. *Addiction, 109*(3), 500-507. doi:10.1111/add.12410

Gonzalez, M. E., & Hoza, C. (1978). Small-area estimation with application to unemployment and housing estimates. *Journal of the American statistical association, 73*(361), 7-15.

Ha, I. D., Lee, Y., & Song, J. k. J. B. (2001). Hierarchical likelihood approach for frailty models. *88*(1), 233-233.

Ha, I. D., Noh, M., & Lee, Y. (2017). H-likelihood approach for joint modeling of longitudinal outcomes and time-to-event data. *Biometrical Journal, 59*(6), 1122-1143.

Hausman, J. A., Newey, W. K., & Powell, J. L. (1995). Nonlinear errors in variables estimation of some Engel curves. *Journal of Econometrics, 65*(1), 205-233.

Henderson, C. R. (1950). *Estimation of genetic parameters.* Paper presented at the Biometrics.

Hobza, T., & Morales, D. J. J. o. o. s. (2016). Empirical best prediction under unit-level logit mixed models. *32*(3), 661-692.

Huang, X., & Wolfe, R. A. (2002). A frailty model for informative censoring. *Biometrics, 58*(3), 510-520.

Institute, N. C. (1998). Smoking and Tobacco Control Monograph 9: Cigars: Health Effects and Trends. Bethesda, MD. Available at https://cancercontrol.cancer.gov/BRP/tcrb/monographs/9/index.html. Accessed on July 16, 2017.

Jamal, A., Gentzke, A., Hu, S. S., Cullen, K. A., Apelberg, B. J., Homa, D. M., & King, B. A. (2017). Tobacco Use Among Middle and High School Students - United States, 2011-2016. *MMWR Morb Mortal Wkly Rep, 66*(23), 597-603. doi:10.15585/mmwr.mm6623a1

Jensen, R. P., Luo, W., Pankow, J. F., Strongin, R. M., & Peyton, D. H. (2015). Hidden formaldehyde in e-cigarette aerosols. *N Engl J Med, 372*(4), 392-394. doi:10.1056/NEJMc1413069

Jiang, J., Lahiri, P., & Wan, S.-M. (2002). A unified jackknife theory for empirical best prediction with M-estimation. *The annals of statistics, 30*(6), 1782-1810.

Jiang, J., & Lahiri, P. J. T. (2006). Mixed model prediction and small area estimation. *15*(1), 1.

Kostygina, G., Glantz, S. A., & Ling, P. M. (2016). Tobacco industry use of flavours to recruit new users of little cigars and cigarillos. *Tob Control, 25*(1), 66-74. doi:10.1136/tobaccocontrol-2014-051830

Laplace, P. S. (1986). Memoir on the probability of the causes of events. *Statistical Science, 1*(3), 364-378.

Lee, Y., Jang, M., & Lee, W. (2011). Prediction interval for disease mapping using hierarchical likelihood. *Computational Statistics, 26*(1), 159-179.

Lee, Y., Nelder, J. A., & Pawitan, Y. (2006). *Generalized linear models with random effects: unified analysis via H-likelihood*: Chapman and Hall/CRC.

Lee, Y., & Nelder, J. A. J. J. o. t. R. S. S. S. B. (1996). Hierarchical generalized linear models. *58*(4), 619-656.

Lee, Y., Ronnegard, L., & Noh, M. (2017). *Data analysis using hierarchical generalized linear models with R*: Chapman and Hall/CRC.

Leventhal, A. M., Strong, D. R., Kirkpatrick, M. G., Unger, J. B., Sussman, S., Riggs, N. R., . . . Audrain-McGovern, J. (2015). Association of Electronic Cigarette Use With Initiation of Combustible Tobacco Product Smoking in Early Adolescence. *JAMA, 314*(7), 700-707. doi:10.1001/jama.2015.8950 2428954 [pii]

MacGibbon, B., & Tomberlin, T. J. (1987). *Small area estimates of proportions via empirical Bayes techniques*: Faculty of Commerce and Administration, Concordia University.

Martuzzi, M., & Elliott, P. (1996). Empirical Bayes estimation of small area prevalence of non-rare conditions. *Stat Med, 15*(17-18), 1867-1873. doi:10.1002/(SICI)1097-0258(19960915)15:17<1867::AID-SIM398>3.0.CO;2-2

Mauro, F., Monleon, V. J., Temesgen, H., & Ford, K. R. J. P. o. (2017). Analysis of area level and unit level models for small area estimation in forest inventories assisted with LiDAR auxiliary information. *12*(12), e0189401.

McCullagh, P. (2018). *Generalized linear models*: Routledge.

McCulloch, C. E., & Searle, S. R. (2004). *Generalized, linear, and mixed models*: John Wiley & Sons.

Molenberghs, G., Verbeke, G., Demétrio, C. G., & Vieira, A. M. J. S. s. (2010). A family of generalized linear models for repeated measures with normal and conjugate random effects. *25*(3), 325-347.

Molina, I., & Marhuenda, Y. J. T. R. J. (2015). sae: An R package for small area estimation. *7*(1), 81-98.

Morris, C. N. (1983). Parametric empirical Bayes inference: theory and applications. *Journal of the American statistical association, 78*(381), 47-55.

National Cancer Institute. (2017). A Socioecological Approach to Addressing Tobacco-Related Health Disparities. National Cancer Institute Tobacco Control Monograph 22. Available at https://cancercontrol.cancer.gov/brp/tcrb/monographs/22/monograph22.html. Accessed on October 20, 2017.

Noh, M., & Lee, Y. (2007). REML estimation for binary data in GLMMs. *Journal of Multivariate Analysis, 98*(5), 896-915. doi:10.1016/j.jmva.2006.11.009

Petrucci, A., & Salvati, N. (2006). Small area estimation for spatial correlation in watershed erosion assessment. *Journal of agricultural, biological, and environmental statistics, 11*(2), 169.

Prasad, N., & Rao, J. (1986). *On the estimation of mean square error of small area predictors.* Paper presented at the Proceedings of the Section on Survey Research Methods, American Statistical Association.

Prasad, N. N., & Rao, J. N. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American statistical association, 85*(409), 163-171.

Primack, B. A., Soneji, S., Stoolmiller, M., Fine, M. J., & Sargent, J. D. (2015). Progression to Traditional Cigarette Smoking After Electronic Cigarette Use Among US Adolescents and Young Adults. *JAMA Pediatr, 169*(11), 1018-1023. doi:10.1001/jamapediatrics.2015.1742

Rahman, A., & Harding, A. (2016). *Small area estimation and microsimulation modeling*: Chapman and Hall/CRC.

Rao, J. N., & Molina, I. (2015). *Small area estimation*: John Wiley & Sons.

Rao, J. N., & Yu, M. (1994). Small-area estimation by combining time-series and cross-sectional data. *Canadian Journal of Statistics, 22*(4), 511-528.

Rigotti, N. A., & Kalkhoran, S. (2017). Reducing Health Disparities by Tackling Tobacco Use. In: Springer.

Rigotti, N. A., & Kalkhoran, S. (2017). Reducing Health Disparities by Tackling Tobacco Use. *J Gen Intern Med, 32*(9), 961-962. doi:10.1007/s11606-017-4098-7

RStudio Team, J. R., Inc., Boston, MA URL https://www.rstudio.com/. (2015). RStudio: integrated development for R. *42*, 14.

Saei, A., & Chambers, R. (2003). Small area estimation under linear and generalized linear mixed models with time and area effects.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics, 6*(2), 461-464.

Searle, S. R., Casella, G., & McCulloch, C. E. (2009). *Variance components* (Vol. 391): John Wiley & Sons.

Shi, C., & Shi, M. C. (2013). Package 'BayesSAE'. Retrieved from https://cran.r-project.org/web/packages/BayesSAE/index.html

Shun, Z., & McCullagh, P. (1995). Laplace approximation of high dimensional integrals. *Journal of the Royal Statistical Society: Series B (Methodological), 57*(4), 749-760.

Soneji, S., Barrington-Trimis, J. L., Wills, T. A., Leventhal, A. M., Unger, J. B., Gibson, L. A., . . . Sargent, J. D. (2017). Association Between Initial Use of e-Cigarettes and Subsequent Cigarette Smoking Among Adolescents and Young Adults: A Systematic Review and Meta-analysis. *JAMA Pediatr, 171*(8), 788-797. doi:10.1001/jamapediatrics.2017.1488

Stroud, T. (1994). Bayesian analysis of binary survey data. *Canadian Journal of Statistics, 22*(1), 33-45.

Taioli, E., & Wynder, E. L. (1991). Effect of the age at which smoking begins on frequency of smoking in adulthood. *N Engl J Med, 325*(13), 968-969. doi:10.1056/NEJM199109263251318

Torabi, M., & Rao, J. J. J. o. M. A. (2014). On small area estimation under a sub-area level model. *127*, 36-55.

Tsiatis, A., & Davidian, M. (2004). An overview of joint modeling of longitudinal and time-to-event data. *Statistica Sinica, 14*, 793-818.

U.S. Department of Health and Human Services. (2012). Preventing Tobacco Use Among Youth and Young Adults: A Report of the Surgeon General. *Atlanta: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health*.

U.S. Department of Health and Human Services. (2014). The Health Consequences of Smoking—50 Years of Progress: A Report of the Surgeon General. Atlanta: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health.

U.S. Department of Health and Human Services. (2016). E-Cigarette Use Among Youth and Young Adults: A Report of the Surgeon General. *Atlanta: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health*.

Walker, K. (2018). Tidycensus: Load us census boundary and attribute data as' tidyverse'and'sf'-ready data frames. In: R package version 0.4.

Wang, C., Hsu, L., Feng, Z., & Prentice, R. L. (1997). Regression calibration in failure time regression. *Biometrics*, 131-145.

Wang, T. W., Gentzke, A., Sharapova, S., Cullen, K. A., Ambrose, B. K., & Jamal, A. (2018). Tobacco Product Use Among Middle and High School Students—United States, 2011–2017. *Morbidity and Mortality Weekly Report, 67*(22), 629.

Wills, T. A., Knight, R., Sargent, J. D., Gibbons, F. X., Pagano, I., & Williams, R. J. (2016). Longitudinal study of e-cigarette use and onset of cigarette smoking among high school students in Hawaii. *Tob Control*. doi:10.1136/tobaccocontrol-2015-052705

Wills, T. A., & Soneji, S. S. (2018). Individual-Level and Ecological Studies. *J Adolesc Health, 62*(5), 507-508. doi:10.1016/j.jadohealth.2018.03.002

Yasui, Y., Liu, H., Benach, J., & Winget, M. (2000). An empirical evaluation of various priors in the empirical Bayes estimation of small area disease risks. *Stat Med, 19*(17-18), 2409-2420. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/10960862

You, Y., & Chapman, B. J. S. M. (2006). Small area estimation using area level models and estimated sampling variances. *32*(1), 97.

You, Y. J. P. o. S. M. S., Statistical Society of Canada. (2008). Small area estimation using area level models with model checking and applications.