Statistica Sinica **26** (2016), 569-587
doi:http://dx.doi.org/10.5705/ss.2013.383

A WEIGHTED COMPOSITE LIKELIHOOD APPROACH FOR ANALYSIS OF SURVEY DATA UNDER TWO-LEVEL MODELS

Grace Y. Yi, J. N. K. Rao and Haocheng Li

University of Waterloo, Carleton University and University of Calgary

Abstract: Multi-level models provide a convenient framework for analyzing data from survey samples with hierarchical structures. Inferential procedures that take account of survey design features are well established for single-level (or marginal) models. However, methods that are valid for general multi-level models are somewhat limited. This paper presents a unified method for two-level models, based on a weighted composite likelihood approach, that takes account of design features and provides valid inferences even for small sample sizes within level 2 units. The proposed method has broad applicability and is straightforward to implement. Empirical studies have demonstrated that the method performs well in estimating the model parameters. Moreover, this research has an important implication: it provides a particular scenario to showcase the unique merit of the composite likelihood method where the likelihood method would not work.

Key words and phrases: Complex sampling design, composite likelihood, design-based inference, multi-level model, super-population model, variance estimation.

1. Introduction

Multi-stage sampling has been widely used in survey studies. For example, education surveys often involve two-stage sampling designs. First-stage sampling units consist of schools which may be selected with probabilities proportional to school size, and second-stage units include students who may be chosen by stratified random sampling from the selected schools. Multi-level models are natural and useful tools to analyze survey data with hierarchical structures. In particular, generalized linear mixed models have been widely employed to accommodate two-stage sampling: a sample of level 2 units (clusters) is selected according to a specified design, and then a sample of elements (or level 1 units) is selected from each sampled level 2 unit according to another specified design. Discussion on multi-stage sampling methods can be found in Cochran (1977), Rao, Wu, and Yue (1992), and Rust and Rao (1996), among others.

When carrying out inference about the model parameters of a multi-level model, it is important to accommodate sampling design features, such as stratification, clustering, and unequal selection probabilities; otherwise, misleading or erroneous results may result (e.g., Pfeffermann et al. (1998); Rao and Roberts (1998); Rao, Verret, and Hidiroglou (2013)). In the case of single-level models, incorporating selection probabilities into inference procedures has been well studied by many authors, including Binder (1983) and Skinner (1989). Although there are some important contributions on multi-level models for survey data (Pfeffermann et al. (1998); Stapleton (2002); Korn and Graubard (2003); Kovacevic and Rai (2003); Grilli and Pratesi (2004); Pfeffermann, Moura, and Silva (2006); Asparouhov (2006); Rabe-Hesketh and Skrondal (2006)), issues in this area remain relatively unresolved. Thus, asymptotic properties of proposed methods are largely unknown, and consistent estimators for general multi-level models are typically not available (Asparouhov (2006)).

We address the problem by exploring a unified inferential procedure for multi-level models featuring survey data with sampling probabilities incorporated. Our method provides valid inferences on model parameters and leads to consistent estimators under a joint model and design setup. Our approach is based on the composite likelihood formulation. The composite likelihood method was initially considered by Besag (1974), and then systematically discussed by Lindsay (1988). This inference strategy has attracted a wide variety of applications, including analysis of longitudinal data (e.g., He and Yi (2011); Yi, Zeng, and Cook (2011); Li and Yi (2013)), spatial data (e.g., Heagerty and Lele (1998)), and image data (e.g., Nott and Rydén (1999)). The use of the composite likelihood, especially pairwise likelihood, has received increasing attention in recent years due to its advantages, including simplicity in defining the objective function, computational advantages when dealing with data with complex structures, and robustness of model specification (Lindsay, Yi, and Sun (2011)). A recent review can be found in Varin (2008) and Varin, Reid, and Firth (2011). Rao, Verret, and Hidiroglou (2013) introduced weighted log pairwise likelihood that can handle general multi-level methods and empirically studied the performance of the method for a simple normal two-level model. Our paper provides extensions and establishes theoretical properties of the method proposed by Rao, Verret, and Hidiroglou (2013).

The remainder of the paper is organized as follows. In Section 2, we introduce notation and the basic framework. General methodology for two-level models is presented in Section 3. Empirical performance of the proposed method is assessed under a simple logistic mixed model in Section 4 and a simple linear mixed model in Section 5. General discussion is in Section 6.

2. Notation and Framework

We consider an inference framework that pertains to two sources of randomness: the probability sampling design for a finite population and the assumed model for a super-population. Under this framework, the finite population is treated as a random sample from the super-population, and the survey sample is regarded as a random sample from the finite population. To be specific, suppose there is a super-population model ξ. Assume that a sequence of finite populations, indexed by ν, is randomly generated from the super-population model ξ, each of size M_ν. For a given ν, a sample of size m_ν is taken from the finite population with index ν according to a specified probability sampling design d. Assume that both M_ν and m_ν tend to infinity as $\nu \to \infty$.

This theoretical framework is useful for understanding and developing subsequent statistical properties. In reality, however, only one finite population and one survey sample from this population are available, so we drop the subscript ν in the discussion. Here we consider a finite population having a two-level structure. Let N be the number of level 2 units in the population and M_i be the number of level 1 units in the level 2 unit i, so that the total number of units in the population is $M = \sum_{i=1}^{N} M_i$. Let Y_{ij} be the response variable for level 1 unit j in level 2 unit i, and $\mathbf{x}_{ij}$ be the associated covariate vector, $i = 1, \ldots, N$, and $j = 1, \ldots, M_i$. Correspondingly, the super-population model from which this finite population is generated is assumed to match the design two-level structure.

Let $\mathbf{x}_i = (\mathbf{x}_{i1}^{\mathrm{T}}, \ldots, \mathbf{x}_{iM_i}^{\mathrm{T}})^{\mathrm{T}}$. In the first step, we assume that given covariate $\mathbf{x}_i$ for level 2 unit i and random effects $\mathbf{u}_i$, the Y_{ij} are independently distributed as

$$Y_{ij} \sim f_{y|u}(y_{ij}|\mathbf{x}_i, \mathbf{u}_i; \boldsymbol{\theta}_y), \quad j = 1, \ldots, M_i, \tag{2.1}$$

where $f_{y|u}$ is a known density function and $\boldsymbol{\theta}_y$ is the associated parameter vector. We assume that $f_{y|u}(y_{ij}|\mathbf{x}_i, \mathbf{u}_i; \boldsymbol{\theta}_y) = f_{y|u}(y_{ij}|\mathbf{x}_{ij}, \mathbf{u}_i; \boldsymbol{\theta}_y)$ for $j = 1, \ldots, M_i$, an assumption that is often made in practice.

In the second step we model random effects by assuming that the $\mathbf{u}_i$ are independently, and marginally distributed as

$$\mathbf{u}_i \ \sim \ f_u(\mathbf{u}_i; \boldsymbol{\theta}_u), \tag{2.2}$$

where $f_u(\mathbf{u}_i; \boldsymbol{\theta}_u)$ is a given density function that is indexed by the parameter $\boldsymbol{\theta}_u$. The population (or census) log likelihood, based on (2.1) and (2.2), can be written as

$$\log L_c(\boldsymbol{\theta}) = \sum_{i=1}^{N} \log L_{ci}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \ell_{ci}(\boldsymbol{\theta}) = \ell_c(\boldsymbol{\theta}), \tag{2.3}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_y^{\mathrm{T}}, \boldsymbol{\theta}_u^{\mathrm{T}})^{\mathrm{T}}$ is the vector of model parameters and

$$L_{ci}(\boldsymbol{\theta}) = \int \exp\Big\{ \sum_{j=1}^{M_i} \log f_{y|u}(y_{ij}|\mathbf{x}_{ij}, \mathbf{u}_i; \boldsymbol{\theta}_y) \Big\} f_u(\mathbf{u}_i; \boldsymbol{\theta}_u) d\mathbf{u}_i. \tag{2.4}$$

This model formulation covers both linear mixed models and generalized linear mixed models.

With informative sampling of level 2 units and of elements within sampled level 2 units, this population model may not hold for the sample. In that case, standard methods for multi-level models that ignore the design and assume (2.1) with (2.2) holding for the sample can lead to asymptotically biased estimators of model parameters $\boldsymbol{\theta}_y$ and $\boldsymbol{\theta}_u$ (Pfeffermann et al. (1998)). To address this issue, properly incorporating the sampling information into the inference becomes critical. In Section 3, we tackle this problem using the weighted composite likelihood framework in order to attain both validity and robustness of results.

3. General Methodology for Two-Level Models

3.1. Overview

Let the sample consist of n level 2 units, denoted s, with m_i level 1 units (elements) from sampled level 2 unit i, denoted $s(i)$, so that the total number of observations in the sample is $m = \sum_{i=1}^{n} m_i$. Let π_i and $\pi_{j|i}$, respectively, denote the level 2 and level 1 inclusion probabilities associated with level 2 unit i and element j within level 2 unit i. Then the level 2 and level 1 design weights are given by $w_i = \pi_i^{-1}$ and $w_{j|i} = \pi_{j|i}^{-1}$, respectively. If the sampling design is not informative, then the population model also holds for the sample, and the resulting sample log likelihood is given by

$$\log L(\boldsymbol{\theta}) = \sum_{i \in s} \log L_i(\boldsymbol{\theta}) = \sum_{i \in s} \ell_i(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}), \tag{3.1}$$

where

$$L_i(\boldsymbol{\theta}) = \int \exp\Big\{ \sum_{j \in s(i)} \log f_{y|u}(y_{ij} | \mathbf{x}_{ij}, \mathbf{u}_i; \boldsymbol{\theta}_y) \Big\} f_u(\mathbf{u}_i; \boldsymbol{\theta}_u) d\mathbf{u}_i. \tag{3.2}$$

Asparouhov (2006) and Rabe-Hesketh and Skrondal (2006) proposed a weighted sample log pseudo-likelihood obtained by inserting the design weights w_i and $w_{j|i}$:

$$\ell_w(\boldsymbol{\theta}) = \sum_{i \in s} \widetilde{w}_i \ell_{wi}(\boldsymbol{\theta}), \tag{3.3}$$

where $\ell_{wi}(\boldsymbol{\theta}) = \log L_{wi}(\boldsymbol{\theta})$, and

$$L_{wi}(\boldsymbol{\theta}) = \int \exp\Big\{ \sum_{j \in s(i)} \widetilde{w}_{j|i} \log f_{y|u}(y_{ij} | \mathbf{x}_{ij}, \mathbf{u}_i; \boldsymbol{\theta}_y) \Big\} f_u(\mathbf{u}_i; \boldsymbol{\theta}_u) d\mathbf{u}_i, \tag{3.4}$$

with normalized weights $\widetilde{w}_i = n w_i / \sum_{i \in s} w_i$, and $\widetilde{w}_{j|i} = m_i w_i / \sum_{j \in s(i)} w_{j|i}$ such that $\sum_{i \in s} \widetilde{w}_i = n$ and $\sum_{j \in s(i)} \widetilde{w}_{j|i} = m_i$.

Maximizing the weighted sample log pseudo-likelihood $\ell_w(\boldsymbol{\theta})$ gives a pseudo maximum likelihood (PML) estimator of $\boldsymbol{\theta}$. This method, however, results in asymptotically biased estimators of level 2 model parameters $\boldsymbol{\theta}_u$. It is noted that the weighted sample log pseudo-likelihood $\ell_w(\boldsymbol{\theta})$ is a design-biased estimator of the "census" log likelihood (2.3). Basically, consistency with respect to both design and model of the PML estimators of variance components in the model requires both the number of sample clusters, n, and the within cluster sample sizes, m_i, to be large (Rao, Verret, and Hidiroglou (2013)).

Under simple random sampling of both level 1 and level 2 units, we have $w_i = N/n$ and $w_{j|i} = M_i/m_i$, so that $\widetilde{w}_i = 1$ and $\widetilde{w}_{j|i} = 1$. Hence, under this noninformative sampling design, $\ell_w(\boldsymbol{\theta})$ reduces to the unweighted log likelihood $\ell(\boldsymbol{\theta})$. In this case, the PML estimates are identical to the customary estimates based on the unweighted log likelihood given by (3.1) and (3.2).

3.2. "Census" composite likelihood

Instead of performing the estimation procedure based on estimating the census full log likelihood, we propose to conduct estimation using the composite likelihood method. Let $L_{ij} = f(y_{ij}|\mathbf{x}_i)$ be the probability density or mass function of Y_{ij}, determined by

$$L_{ij} = \int f_{y|u}(y_{ij}|\mathbf{x}_{ij}, \mathbf{u}_i) f_u(\mathbf{u}_i) d\mathbf{u}_i.$$

For $j \neq k$, let $L_{ijk} = f(y_{ij}, y_{ik}|\mathbf{x}_i)$ be the joint probability density or mass function for paired responses (Y_{ij}, Y_{ik}); this is determined by

$$L_{ijk} = \int f_{y|u}(y_{ij}|\mathbf{x}_{ij}, \mathbf{u}_i) f_{y|u}(y_{ik}|\mathbf{x}_{ik}, \mathbf{u}_i) f_u(\mathbf{u}_i) d\mathbf{u}_i.$$

The dependence on the parameter $\boldsymbol{\theta}$ is suppressed in the notation.

A "census" composite likelihood can be formulated based on the marginal pairwise distributions,

$$C(\boldsymbol{\theta}) = \prod_{i=1}^{N} \prod_{j<k} L_{ijk}^{B_{jk}} L_{ij}^{B_j} L_{ik}^{B_k},$$

where B_{jk}, B_j, and B_k are weights that can be user-specified to enhance efficiency or to facilitate some specific features of the formulation. For instance, letting $B_{jk} = 1$ and $B_j = B_k = 0$ leads to the pairwise likelihood $\prod_{i=1}^{N} \prod_{j<k} f(y_{ij}, y_{ik}|\mathbf{x}_i)$; setting $B_{jk} = 2$ and $B_j = B_k = -1$ leads to the conditional pairwise likelihood $\prod_{i=1}^{N} \prod_{j<k} f(y_{ij}|y_{ik}, \mathbf{x}_i) f(y_{ik}|y_{ij}, \mathbf{x}_i)$. Taking $B_{jk} = 0$ and $B_j = B_k = 1$ results in the product of marginal distributions $\prod_{i=1}^{N} \prod_{j=1}^{M_i} f(y_{ij}|\mathbf{x}_i)$ with possible association among response components ignored, and setting $B_{jk} = 1$ and $B_j = B_k =$

-1 yields the so-called Hoeffding formulation $\prod_{i=1}^{N}\prod_{j<k}[f(y_{ij},y_{ik}|\mathbf{x}_i)/\{f(y_{ij}|\mathbf{x}_i)$ $f(y_{ik}|\mathbf{x}_i)\}]$ (Lindsay, Yi, and Sun (2011)). Some discussion on choosing weights is given by Joe and Lee (2009), and Lindsay, Yi, and Sun (2011).

3.3. Point estimation

A "census" log pairwise likelihood under the assumed two-level model given by (2.1) and (2.2) is obtained as

$$\ell_c(\boldsymbol{\theta}) = \sum_{i=1}^{N}\sum_{1\le j<k\le M_i}(B_{jk}\ell_{ijk} + B_j\ell_{ij} + B_k\ell_{ik}), \tag{3.5}$$

where $\ell_{ij} = \log L_{ij}$, and $\ell_{ijk} = \log L_{ijk}$. Here we consider the case with $B_j = 0$ and $B_k = 0$. Extensions to accommodating other weights are discussed later. We consider a "census" log all-pairwise likelihood

$$\ell_c(\boldsymbol{\theta}) = \sum_{i=1}^{N}\sum_{1\le j<k\le M_i} B_{jk}\ell_{ijk}.$$

Using the within-cluster joint inclusion probabilities, $\pi_{jk|i}$, we consider a weighted "sample" log all-pairwise likelihood

$$\ell_{wc}(\boldsymbol{\theta}) = \sum_{i\in s} w_i \sum_{j<k,j,k\in s(i)} w_{jk|i}B_{jk}\ell_{ijk}, \tag{3.6}$$

where $w_{jk|i} = \pi_{jk|i}^{-1}$.

There is an important difference in the pairwise likelihood and the full likelihood formulations. In the formulation of the weighted "sample" version corresponding to the full likelihood formulation (3.3) and (3.4) in Section 3.1, the two-stage sampling weights appear in a non-linear form. Therefore, the design-based expectation of the weighted sample version $\ell_w(\boldsymbol{\theta})$ cannot recover the census full log likelihood $\ell(\boldsymbol{\theta})$. However, when using the "census" log pairwise likelihood (3.5), the two-stage sampling weights enter (3.5) in a linear fashion to form a weighted "sample" version (3.6), hence $\ell_{wc}(\boldsymbol{\theta})$ is design unbiased for the census $\ell_c(\boldsymbol{\theta})$.

Solving

$$\mathbf{U}_{wc}(\boldsymbol{\theta}) = \frac{\partial \ell_{wc}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i\in s} w_i \mathbf{U}_{iwc}(\boldsymbol{\theta}) = \mathbf{0} \tag{3.7}$$

for $\boldsymbol{\theta}$ leads to the weighted composite likelihood estimator, $\widehat{\boldsymbol{\theta}}_w$, of $\boldsymbol{\theta}$, where

$$\mathbf{U}_{iwc}(\boldsymbol{\theta}) = \sum_{j<k,j,k\in s(i)} w_{jk|i}B_{jk}\mathbf{s}_{ijk}$$

with $\mathbf{s}_{ijk} = \partial \ell_{ijk}/\partial \boldsymbol{\theta}$.

One notices from (3.6) and (3.7) that to implement the proposed method, we need not only the inclusion probabilities $\pi_i = w_i^{-1}$ but also the within level 2 unit joint inclusion probabilities $\pi_{jk|i} = w_{jk|i}^{-1}$. This information is often available in such settings as simple random or stratified random sampling within level 2 units, or when the within level 2 unit sampling fraction is small. If such information is not available, one may employ an approximation to $\pi_{jk|i}$. When sampling within level 2 units is based on unequal probability sampling, then approximations to $\pi_{jk|i}$ depending only on the marginal inclusion probabilities $\pi_{j|i}$ can be utilized; see Haziza, Mecatti, and Rao (2008) for details.

Now we show that $E_\xi E_d\{\mathbf{U}_{wc}(\boldsymbol{\theta})\} = 0$, where E_ξ and E_d stand for the expectation taken with respect to model ξ and sampling design d, respectively. By the nature of the two-stage design weights $w_{j|i}$ and w_i, the inner expectation $E_d\{\mathbf{U}_{wc}(\boldsymbol{\theta})\}$ recovers the census composite score function, $\mathbf{U}_c(\boldsymbol{\theta}) = (\partial/\partial\boldsymbol{\theta})\{\ell_c(\boldsymbol{\theta})\}$. Then the unbiasedness of the latter function ensures zero expectation of the weighted composite score function taken with respect to the design and the model. As a result, the weighted composite likelihood estimator $\widehat{\boldsymbol{\theta}}_w$ is consistent from the perspective of the joint model and design. In particular, $\widehat{\boldsymbol{\theta}}_w$ is design-model consistent for $\boldsymbol{\theta}$ as the number n of level 2 units in the sample approaches ∞, even when the within-level 2 unit sizes, m_i, are small. In Appendix B, we prove the proof of the following theorem.

Theorem 1. *Under regularity conditions stated in Appendix A,*

$$\widehat{\boldsymbol{\theta}}_w \xrightarrow{p} \boldsymbol{\theta} \text{ as } n \to \infty,$$

where "p" denotes convergence in probability with respect to joint model ξ and sampling design d.

3.4. Variance estimation

To evaluate the precision of the estimator $\widehat{\boldsymbol{\theta}}_w$, we need to accommodate two types of variability induced from sampling. The first type of variability arises from a census fit to the super-population model using data of an entire finite population, while the other comes from using observations of a sample taken from this finite population according to a given sampling design. To be precise, the covariance matrix of the estimator $\widehat{\boldsymbol{\theta}}_w$ is given by

$$\text{cov}_{\xi d}(\widehat{\boldsymbol{\theta}}_w) = \text{cov}_\xi\{E_d(\widehat{\boldsymbol{\theta}}_w)\} + E_\xi\{\text{cov}_d(\widehat{\boldsymbol{\theta}}_w)\}.$$

If $\boldsymbol{\theta}_U = E_d(\widehat{\boldsymbol{\theta}}_w)$, then $\boldsymbol{\theta}_U$ can be viewed as a finite population (or census) quantity that is unbiasedly estimated by the estimator $\widehat{\boldsymbol{\theta}}_w$. As discussed by

Demnati and Rao (2010), and Carrillo, Chen, and Wu (2010), if $\mathrm{cov}_\xi(\boldsymbol{\theta}_U)$ has the order of $1/N$ and the sampling fraction n/N is small, then we approximate $\mathrm{cov}_{\xi d}(\widehat{\boldsymbol{\theta}}_w)$ with $E_\xi\{\mathrm{cov}_d(\widehat{\boldsymbol{\theta}}_w)\}$,

$$\mathrm{cov}_{\xi d}(\widehat{\boldsymbol{\theta}}_w) \approx E_\xi\{\mathrm{cov}_d(\widehat{\boldsymbol{\theta}}_w)\}, \tag{3.8}$$

which suggests that an estimator of $\mathrm{cov}_d(\widehat{\boldsymbol{\theta}}_w)$ can be approximately taken as a design-model based estimator of the covariance matrix $\mathrm{cov}_{\xi d}(\widehat{\boldsymbol{\theta}}_w)$.

We discuss an approach to estimate the design-based covariance $\mathrm{cov}_d(\widehat{\boldsymbol{\theta}}_w)$ using a Taylor series expansion, similar to Binder (1983). Let $\boldsymbol{\theta}_N$ denote the solution to the census composite score equation $\mathbf{U}_c(\boldsymbol{\theta}) = \partial\ell_c(\boldsymbol{\theta})/\partial\boldsymbol{\theta} = \mathbf{0}$; we call $\boldsymbol{\theta}_N$ the census parameter. Noting that $\mathbf{U}_{wc}(\widehat{\boldsymbol{\theta}}_w) = \mathbf{0}$ and expanding $\mathbf{U}_{wc}(\widehat{\boldsymbol{\theta}}_w) = \mathbf{0}$ at $\boldsymbol{\theta} = \boldsymbol{\theta}_N$ by a Taylor series expansion leads to

$$\begin{aligned}\mathbf{0} &= \mathbf{U}_{wc}(\widehat{\boldsymbol{\theta}}_w)\\ &= \mathbf{U}_{wc}(\boldsymbol{\theta}_N) + \frac{\partial \mathbf{U}_{wc}(\boldsymbol{\theta}_N)}{\partial \boldsymbol{\theta}^{\mathrm{T}}}(\widehat{\boldsymbol{\theta}}_w - \boldsymbol{\theta}_N) + o_p(\frac{1}{\sqrt{n}}).\end{aligned}$$

If $\boldsymbol{\Gamma}_{wc}(\boldsymbol{\theta}_N) = -\partial\mathbf{U}_{wc}(\boldsymbol{\theta}_N)/\partial\boldsymbol{\theta}^{\mathrm{T}}$, then by $\mathbf{U}_{wc}(\boldsymbol{\theta}_N) = \mathbf{U}_c(\boldsymbol{\theta}_N) + O_p(N/\sqrt{n})$, we obtain

$$\begin{aligned}\widehat{\boldsymbol{\theta}}_w - \boldsymbol{\theta}_N &= \{\boldsymbol{\Gamma}_{wc}(\boldsymbol{\theta}_N)\}^{-1}\mathbf{U}_{wc}(\boldsymbol{\theta}_N) + o_p(\frac{1}{\sqrt{n}})\\ &= \{\boldsymbol{\Gamma}_{c}(\boldsymbol{\theta}_N)\}^{-1}\mathbf{U}_{wc}(\boldsymbol{\theta}_N) + o_p(\frac{1}{\sqrt{n}}),\end{aligned}$$

where the finite population quantity $\boldsymbol{\Gamma}_c(\boldsymbol{\theta}_N) = -\partial\mathbf{U}_c(\boldsymbol{\theta}_N)/\partial\boldsymbol{\theta}^{\mathrm{T}}$ is used to replace the sample quantity $\boldsymbol{\Gamma}_{wc}(\boldsymbol{\theta}_N)$. As a result, we obtain

$$\mathrm{cov}_d(\widehat{\boldsymbol{\theta}}_w) \approx \{\boldsymbol{\Gamma}_c(\boldsymbol{\theta}_N)\}^{-1}\mathrm{cov}_d\{\mathbf{U}_{wc}(\boldsymbol{\theta}_N)\}\{\boldsymbol{\Gamma}_c(\boldsymbol{\theta}_N)\}^{-1\mathrm{T}}. \tag{3.9}$$

The middle term $\mathrm{cov}_d\{\mathbf{U}_{wc}(\boldsymbol{\theta}_N)\}$ in (3.9) is viewed as a finite population quantity that incorporates sampling design features, and can be expressed in terms of the between- and within-level 2 units variabilities associated with the sampling design. We write

$$\begin{aligned}\mathrm{cov}_d\{\mathbf{U}_{wc}(\boldsymbol{\theta}_N)\} &= \mathrm{cov}_d\Big\{\sum_{i\in s} w_i E_d(\mathbf{U}_{iwc}|s)\Big\} + E_d\Big\{\mathrm{cov}_d(\sum_{i\in s} w_i \mathbf{U}_{iwc}|s)\Big\}\\ &= \mathrm{cov}_d\Big(\sum_{i\in s} w_i \mathbf{U}_i\Big) + E_d\Big\{\sum_{i\in s} w_i^2 \mathrm{cov}_d(\mathbf{U}_{iwc}|s)\Big\},\end{aligned} \tag{3.10}$$

where $\mathbf{U}_i = \mathbf{U}_i(\boldsymbol{\theta}_N) = \sum_{j<k} B_{jk}\mathbf{s}_{ijk}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_N}$ is the "census" version corresponding to the sample version $\mathbf{U}_{iwc} = \mathbf{U}_{iwc}(\boldsymbol{\theta}_N)$.

To evaluate the second term $E_d\{\sum_{i\in s} w_i^2 \text{cov}_d(\mathbf{U}_{iwc}|s)\}$, we define an inclusion indicator variable

$$R_{ij} = \begin{cases} 1, & \text{if } i \in s, j \in s(i), \\ 0, & \text{otherwise,} \end{cases}$$

so that $E_d\{R_{ij}R_{ik}w_{jk|i}|s \text{ is chosen}\} = 1$. Then, conditional on that $i \in s$, we write

$$\begin{aligned} \text{cov}_d(\mathbf{U}_{iwc}) &= \text{cov}_d\Big\{ \sum_{j<k,j,k\in s(i)} w_{jk|i}B_{jk}\mathbf{s}_{ijk} \Big\} \\ &= \text{cov}_d\Big\{ \sum_{1\le j<k\le M_i} R_{ij}R_{ik}(w_{jk|i}B_{jk}\mathbf{s}_{ijk}) \Big\}. \end{aligned}$$

As a result, to precisely calculate $\text{cov}_d(\mathbf{U}_{iwc})$, the level 1 inclusion probabilities of quadruples (j,k,s,t) are needed, which is often feasible in practice.

Precise evaluation of (3.10) leads to difficulties of requiring fourth order within-level 2 unit inclusion probabilities, and hence the estimation of (3.10) is also complex. We follow the customary practice of treating the sample level 2 units as if they were selected with replacement with probabilities p_i, where p_i is a size measure and $\pi_i = np_i$. For example, the Rao-Sampford method of unequal probability sampling ensures that $\pi_i = np_i$ (Rao (1965); Sampford (1967)). As a result, we write

$$\mathbf{U}_{wc}(\boldsymbol{\theta}) = n^{-1}\sum_{i\in s}\widetilde{\mathbf{U}}_{iwc}(\boldsymbol{\theta}),$$

where $\widetilde{\mathbf{U}}_{iwc}(\boldsymbol{\theta}) = \mathbf{U}_{iwc}(\boldsymbol{\theta})/p_i$ are independent with the same mean and the same variance from the design perspective, and $\mathbf{U}_{iwc}(\boldsymbol{\theta}) = \sum_{j<k,j,k\in s(i)} w_{jk|i}B_{jk}\mathbf{s}_{ijk}(\boldsymbol{\theta})$.

Consequently, we estimate $\text{cov}_d\{\mathbf{U}_{wc}(\boldsymbol{\theta}_N)\}$ as

$$\widehat{\text{cov}}_d\{\mathbf{U}_{wc}(\boldsymbol{\theta}_N)\} = \{n(n-1)\}^{-1}\sum_{i\in s}(\widetilde{\mathbf{U}}_{iwc} - \mathbf{U}_{wc})(\widetilde{\mathbf{U}}_{iwc} - \mathbf{U}_{wc})^{\mathrm{T}}$$

evaluated at the estimator $\widehat{\boldsymbol{\theta}}_w$. As $\mathbf{U}_{wc}(\widehat{\boldsymbol{\theta}}_w)$ is zero, we then obtain

$$\begin{aligned} \widehat{\text{cov}}_d\{\mathbf{U}_{wc}(\boldsymbol{\theta}_N)\} &= \{n(n-1)\}^{-1}\sum_{i\in s}(\widetilde{\mathbf{U}}_{iwc} - \mathbf{U}_{wc})(\widetilde{\mathbf{U}}_{iwc} - \mathbf{U}_{wc})^{\mathrm{T}} \\ &= \{n(n-1)\}^{-1}\sum_{i\in s}\widetilde{\mathbf{U}}_{iwc}\widetilde{\mathbf{U}}_{iwc}^{\mathrm{T}} \\ &= \{n(n-1)\}^{-1}\sum_{i\in s}\Big(\frac{\mathbf{U}_{iwc}}{p_i}\Big)\Big(\frac{\mathbf{U}_{iwc}^{\mathrm{T}}}{p_i}\Big) \\ &= \frac{n}{(n-1)}\sum_{i\in s} w_i^2\mathbf{U}_{iwc}\mathbf{U}_{iwc}^{\mathrm{T}}, \end{aligned} \tag{3.11}$$

where $w_i = 1/\pi_i$ and $\mathbf{U}_{iwc} = \mathbf{U}_{iwc}(\widehat{\boldsymbol{\theta}}_w)$. It now follows from (3.8), (3.9) and (3.11) that an approximate estimator of $\mathrm{cov}_{\xi d}(\widehat{\boldsymbol{\theta}}_w)$ is given by

$$\widehat{\mathrm{cov}}_{\xi d}(\widehat{\boldsymbol{\theta}}_w) = \{\Gamma_{wc}(\widehat{\boldsymbol{\theta}}_w)\}^{-1}\widehat{\mathrm{cov}}_d\{\mathbf{U}_{wc}(\boldsymbol{\theta}_N)\}\{\Gamma_{wc}(\widehat{\boldsymbol{\theta}}_w)\}^{-1\mathrm{T}}. \tag{3.12}$$

The estimator (3.12) should perform well in estimating $\mathrm{cov}_{\xi d}(\widehat{\boldsymbol{\theta}}_w)$ if n/N is sufficiently small.

4. Simulation Study: Logistic Mixed Model

In this section we report the results of a simulation study to assess the performance of the proposed method under the logistic mixed model. Let Y_{ij} be a binary response with conditional mean $\mu_{ij} = E(Y_{ij}|\mathbf{x}_i, u_i)$, given covariate $\mathbf{x}_i$ and random effects u_i. Response measurements are generated from the model

$$\text{logit } (\mu_{ij}) = \beta_0 + \beta_1 X_{ij} + u_i,$$

where the X_{ij} are generated from the standard normal distribution and held fixed over simulation runs, the u_i are independent and identically distributed (i.i.d.) as $N(0, \sigma^2)$ with $\sigma^2 = 3$, and we let $\beta_0 = 0.5$ and $\beta_1 = 3$. We evaluated the performance of the proposed method under two different sampling strategies: (A). $N = n = 400$, and (B). $N = 4{,}000$, $n = 400$. In Case A, level 2 units are equivalent to strata because all level 2 units are sampled. In Case B, level 2 units are selected via simple random sampling.

We selected $m_i = m = 5$ level 1 units from each level 2 unit, using the Rao-Sampford method of sampling without replacement with probability proportional to specified size measures z_{ij} (Pfeffermann, Moura, and Silva (2006); Sampford (1967)). Following Asparouhov (2006), we took

$$z_{ij} = \begin{cases} \delta_B & \text{if} \quad Y_{ij} = 1, \\ 1 & \text{if} \quad Y_{ij} = 0, \end{cases} \tag{4.1}$$

where $0 < \delta_B \leq 1$. This selection mechanism over samples zero outcomes at a rate of $1/\delta_B$. When $\delta_B = 1/2$ and $\delta_B = 1/3$, the over-sampling rates are 2 and 3, respectively. When $\delta_B = 1$, one and zero outcomes are sampled with equal probabilities, so that $\widetilde{w}_{j|i} = 1$.

We used the design-model approach to simulate $R = 2{,}500$ samples for each specified δ_B. In particular, we generated a population for Case A with $N = n = 400$ and $M_i = 100$ for $i = 1, \ldots, N$ from the model, then selected all the level 2 units and a sample of $m_i = 5$ units from each sampled level 2 unit using the Rao-Sampford method with size measures z_{ij}. To be specific, letting $z_{i+} = \sum_{j=1}^{M_i} z_{ij}$ gives the inclusion probability $\pi_{j|i} = m_i(z_{ij}/z_{i+})$. The pairwise

Table 1. Simulation results for estimation of β_0 under the logistic mixed effect model in Case B.

δ_B	UML*				PML			
	BR**	RRMSE	100AVE	100MSE***	BR	RRMSE	100AVE	100MSE
1/3	-958.40	237.69	1.46	141.30	-273.07	88.87	2.24	19.75
1/2	-635.88	150.00	1.33	56.28	-169.21	48.64	1.51	5.92
1	2.46	24.08	1.35	1.45	2.46	24.08	1.35	1.45

δ_B	WPL			
	BR	RRMSE	100AVE	100MSE
1/3	-1.39	25.80	1.61	1.67
1/2	0.68	23.92	1.44	1.43
1	2.57	24.27	1.37	1.47

$*$: UML denotes the unweighted maximum likelihood method; PML denotes the pseudo maximum likelihood approach; WPL denotes our weighted pairwise likelihood method.
$**$: BR and RRMSE represent bias ratio (%) and relative root mean square error, respectively.
$***$: 100AVE and 100MSE represent 100 times of average variance estimates and mean square error, respectively.

inclusion probability $\pi_{jk|i}$ is calculated using the R package *pps*. For Case B, we generated a population with $N = 4{,}000$ and $M_i = 100$ for $i = 1, \ldots, N$ from the model, and selected $n = 400$ level 2 units by simple random sampling and $m_i = 5$ level 1 units from each level 2 unit in the same manner as in Case A.

We studied the weighted pairwise likelihood method (WPL) with sampling weights taken into account as described in Section 3, where B_{jk} is set as 1. As the integrals involved in pairwise likelihood functions L_{ijk} do not have closed forms for the logistic mixed model, we used adaptive Gaussian quadratures to approximate the integrals with 9 quadrature points (Molenberghs and Verbeke (2005)). As a comparison, we also studied the PML method using the log likelihood (3.3) in Section 3.1. We discussed the customary approach based on maximizing the unweighted log likelihood (labeled as UML) in (3.1).

We report the simulation results in terms of bias ratio (BR), defined as bias/(square root of variance), and relative root mean square error (RRMSE), defined as (square root of mean square error)/(true parameter value). In Case B, the level 2 sampling fraction is not large and hence we approximate the variance estimator using (3.12). We report the average of the approximate variance estimates (AVE) over simulation runs as well as the mean square error of the estimators (MSE).

Simulation results for Case B are reported for β_0 and σ in Tables 1 and 2. The results for β_1 are reported in Section 3.1 of the web-appendix. The differences among the PML method, the UML method, and the WPL method are striking. The proposed WPL method generally outperforms the PML and UML methods

Table 2. Simulation results for estimation of σ under the logistic mixed effect model in Case B.

δ_B	UML*				PML			
	BR**	RRMSE	100AVE	100MSE***	BR	RRMSE	100AVE	100MSE
1/3	11.96	9.57	2.71	2.75	304.21	36.38	3.83	39.72
1/2	5.62	8.89	2.47	2.37	115.95	14.72	2.91	6.51
1	-0.14	8.89	2.39	2.37	-0.14	8.89	2.39	2.37

δ_B	WPL			
	BR	RRMSE	100AVE	100MSE
1/3	-1.14	10.32	3.15	3.19
1/2	0.89	9.32	2.73	2.61
1	0.38	9.19	2.54	2.54

$*$: UML denotes the unweighted maximum likelihood method; PML denotes the pseudo maximum likelihood approach; WPL denotes our weighted pairwise likelihood method.
$**$: BR and RRMSE represent bias ratio (%) and relative root mean square error, respectively.
$***$: 100AVE and 100MSE represent 100 times of average variance estimates and mean square error, respectively.

under informative sampling ($\delta_B < 1$). On the other hand, under noninformative sampling ($\delta_B = 1$), the PML method is identical to the "optimal" UML method based on the customary loglikelihood (3.1), and hence it performs well in terms of BR and RRMSE. However, the WPL method (which reduces to the customary pairwise likelihood method) also performs well but exhibits a small increase in RRMSE relative to the PML and UML methods, as expected. Results on the estimators for Case A, not reported here, exhibit a similar pattern.

Turning to the performance of the variance estimator (3.12) for the WPL method, Table 2 shows that it tracks the MSE for all δ_B. On the other hand, the variance estimators for the PML method (Rabe-Hesketh and Skrondal (2006)) and the UML method perform very poorly for $\delta_B < 1$ because they lead to severe underestimation. For $\delta_B = 1$, the PML and UML methods perform well in tracking the MSE, as expected.

5. Simulation Study: Linear Mixed Model

5.1. Skew normal random effects

Rao, Verret, and Hidiroglou (2013) studied the performance of the WPL method under a nested error linear regression method $Y_{ij} = \beta_0 + X_{ij}\beta_1 + u_i + e_{ij}$ with normally distributed level 2 random effects u_i and random errors e_{ij}. In this section, we relax the normality assumption on u_i by using a skew-normal (SN) family which includes a wide variety of skewed distributions as well as the normal.

We consider the model given by

$$Y_{ij} = \beta_0 + X_{ij}\beta_1 + u_i + e_{ij};\ e_{ij} \sim_{\text{iid}} N(0, \sigma_e^2), \tag{5.1}$$

where the random level 2 effects u_i are assumed to be i.i.d and have a common skew normal distribution $SN(0, \sigma_u^2, \alpha)$ with density function $2\phi(u_i; \sigma_u^2)\Phi(\alpha u_i/\sigma_u)$. Here α is a skewness parameter, $\phi(.; \sigma_u^2)$ is the normal density with zero mean and variance σ_u^2, and $\Phi(t) = \int_{-\infty}^{t} \phi(u; 1)du$ is the $N(0, 1)$ distribution function. This model fits into our two-level model setup with dependence on covariates included. Its more general form was discussed by Lin and Lee (2008). In implementing model (5.1), X_{ij} is generated from $N(0, 1)$ and held fixed over simulation runs.

For inference connected to a skew-normal distribution, Azzalini and Capitanio (1999) and Arellano-Valle and Azzalini (2008) pointed out that singularity arises in the information matrix when the skewness parameter approaches 0, thus breaking down estimation procedures. As a remedy, they suggested to adopt the so-called centered parameterization.

In our simulation studies, we specifically employed the following reparameterization:

$$\theta_1 = \beta_0 + \sqrt{\frac{2}{\pi}}\sigma_u \frac{\alpha}{\sqrt{(1+\alpha^2)}}, \quad \beta_1 = \beta_1, \quad \sigma_e^2 = \sigma_e^2,$$

$$\theta_2 = \sigma_u^2(1 - \frac{2\alpha^2}{\pi(1+\alpha^2)}), \quad \text{and } \gamma_1 = \frac{4-\pi}{2}\frac{(\sqrt{2/\pi}\alpha)^3}{[1+(1-2/\pi)\alpha^2]^{3/2}},$$

leading to a one-to-one correspondence between $(\beta_0, \beta_1, \sigma_e^2, \sigma_u^2, \alpha)$ and $(\theta_1, \beta_1, \sigma_e^2, \theta_2, \gamma_1)$. With this reparameterization, the Newton-Raphson procedure is implemented to obtain the maximum likelihood estimates of the model parameters.

5.2. Simulation setup and results

We conducted a simulation study to evaluate the performance of the proposed method under the skew normal model. Analogous to Section 4, we used the Rao-Sampford method to select level 1 units from each level 2 unit, but size measures z_{ij} are specified differently. Following Asparouhov (2006) and Rao, Verret, and Hidiroglou (2013), we considered both invariant and non-invariant selections as defined below. For invariant selection, we take

$$z_{ij}^{-1} = 1 + \exp\Big[-0.5\Big\{\frac{e_{ij}}{\delta_C} + e_{ij}^*(1-\delta_C^{-2})^{1/2}\Big\}\Big],$$

where e_{ij}^* is independent of e_{ij} but with the same distribution, $N(0, \sigma_e^2)$. In the case of non-invariant selection, we replaced e_{ij} and e_{ij}^* in z_{ij} by u_i+e_{ij} and $u_i^*+e_{ij}^*$,

Table 3. Simulation results for estimation of θ_1 under the linear mixed effect regression model in Case B.

Invariant								
δ_C	UML*				PML			
	BR**	RRMSE	100AVE	100MSE***	BR	RRMSE	100AVE	100MSE
1	410.21	23.61	1.16	20.71	103.49	8.38	1.25	2.61
2	203.14	12.65	1.18	5.94	54.28	6.48	1.24	1.56
3	134.16	9.34	1.18	3.24	37.05	6.08	1.23	1.37
∞	-4.56	5.61	1.18	1.17	2.50	5.70	1.23	1.21
δ_C	WPL							
	BR	RRMSE	100AVE	100MSE				
1	-2.35	6.15	1.40	1.40				
2	-3.42	5.99	1.34	1.33				
3	-2.56	5.94	1.33	1.31				
∞	-4.59	5.91	1.32	1.30				
Non-invariant								
δ_C	UML				PML			
	BR	RRMSE	100AVE	100MSE	BR	RRMSE	100AVE	100MSE
1	323.62	17.63	1.01	11.54	79.95	7.24	1.19	1.95
2	133.04	9.22	1.15	3.16	39.47	6.13	1.21	1.40
3	91.08	7.61	1.17	2.15	29.88	6.01	1.22	1.34
∞	1.53	5.58	1.18	1.16	8.08	5.70	1.22	1.21
δ_C	WPL							
	BR	RRMSE	100AVE	100MSE				
1	3.99	6.13	1.41	1.40				
2	3.22	5.87	1.27	1.28				
3	2.84	5.92	1.26	1.30				
∞	1.41	5.76	1.26	1.23				

$*$: UML denotes the unweighted maximum likelihood method; PML denotes the pseudo maximum likelihood approach; WPL denotes our weighted pairwise likelihood method.
$**$: BR and RRMSE represent bias ratio (%) and relative root mean square error, respectively.
$***$: 100AVE and 100MSE represent 100 times of average variance estimates and mean square error, respectively.

respectively, where u_i^* is independent of u_i and has the distribution $SN(0, \sigma_u^2, \alpha)$. Here we took the parameter settings $\beta_0 = 0.5, \beta_1 = 1, \sigma_e^2 = 2, \sigma_u^2 = 4$, and $\alpha = 2$, which are equivalent to the reparameterization settings $\theta_1 = 1.927, \beta_1 = 1, \sigma_e^2 = 2, \theta_2 = 1.963$, and $\gamma_1 = 0.454$. We considered four values of δ_C: $\delta_C = 1, 2, 3, \infty$, where $\delta_C = \infty$ corresponds to non-informative sampling within each level 2 unit, $\delta_C = 1$ corresponds to most informative sampling, and informativeness decreases as δ_C increases.

Similar to the simulation study in Section 4, we used the design-model approach to simulate R = 2,500 samples for each specified δ_C and separately for invariant and non-invariant selections. In particular, we generated a population for Case A with $N = n = 200$ and $M_i = 100$ for $i = 1, \ldots, N$ from the model and then selected a sample of $m_i = 5$ units from each level 2 unit using the Rao-Sampford method with size measures z_{ij}. For Case B, we generated a population with $N = 2{,}000$ and $M_i = 100$ for $i = 1, \ldots, N$ from the model and selected $n = 200$ level 2 units by simple random sampling and $m_i = 5$ level 1 units from each level 2 unit in the same as in Case A.

We report the results for Case B on the estimator of θ_1 in Table 3. The results for other estimators are reported in Section 3.2 of the web-appendix. It is clearly seen that the PML and UML methods generally produce considerably biased results under informative sampling. On the other hand, the WPL method performs well, leading to reasonably small bias ratios in absolute value under informative sampling. Results for Case A, not reported here, show similar patterns.

6. Discussion

Multi-level models provide a conceptually convenient tool to analyze data arising from complex surveys. In making inferences about model parameters it is important to properly incorporate selection probabilities into inferential procedures. In this paper, we present a general method using the composite likelihood formulation to handle two-level models with survey information accounted for. The proposed estimator is design-model consistent. We applied the proposed method to linear mixed models and logistic mixed models to assess the performance under a variety of circumstances. Our empirical studies demonstrate that biased results would arise if sampling features are ignored, but the proposed weighted composite likelihood method effectively captures the design features. Our method provides good results even when the sample sizes within sampled clusters are small, unlike the PML method.

Our work bridges survey sampling and composite likelihood inference. On the one hand, the composite likelihood method supplies us an effective tool to tackle unsolved problems in survey sampling. With general multi-level models, theoretical results remain largely unexplored, and consistent estimators are typically unavailable. However, by employing the composite likelihood method, in particular, the pairwise likelihood formulation, we are able to work out estimators of the model parameters that are design-model consistent. As well, this survey sampling problem provides a rather unique setting to showcase the merits of the pairwise likelihood formulation, in contrast to the likelihood method. In various existing applications, the composite likelihood method is mostly used as a tool to ease either modeling complexities or computation burdens when the likelihood is not easily obtained or unavailable. Our development uncovers an interesting scenario where the weighted likelihood method breaks down but the weighted composite likelihood method is essential for inference from survey data.

It would be interesting to establish the asymptotic normality of our weighted pairwise likelihood estimator. In principle, this can be established by adapting the theorem of Chen and Rao (2007) in conjunction with the result of Bickel and Freedman (1984) for stratified sampling. This is our future work, in addition to various extensions, including weighted pairwise likelihood ratio tests.

Supplementary Materials

The online supplementary materials include a detailed proof of Theorem 1 and additional simulation results. For completeness, regularity conditions listed in Appendix A are also presented.

Acknowledgements

The authors thank Laura Dumitrescu and the review team for their helpful comments on the initial version of this paper. This research was supported by the Natural Sciences and Engineering Research Council of Canada.

Appendix A: Regularity Conditions

If $\boldsymbol{\psi}_{ijk}(\boldsymbol{\theta}; y_{ij}, y_{ik}) = B_{jk}\mathbf{s}_{ijk}$, then $\mathbf{U}_w(\boldsymbol{\theta}) = \sum_{i\in s} w_i \sum_{j<k,j,k\in s(i)} w_{jk|i}$ $\boldsymbol{\psi}_{ijk}(\boldsymbol{\theta}; y_{ij}, y_{ik})$, $\boldsymbol{\theta} \in \Theta \subset R^p$, where p is the dimension of $\boldsymbol{\theta}$. Let $\boldsymbol{\theta}_0$ be the value such that $E_\xi E_d\{\mathbf{U}_w(\boldsymbol{\theta}_0)\} = \mathbf{0}$, and $h_{ijk}(y_{ij}, y_{ik}) = \sup_{\boldsymbol{\theta}\in\Theta} ||\boldsymbol{\psi}_{ijk}(\boldsymbol{\theta}; y_{ij}, y_{ik})||$ for the triples (i, j, k), where $||\cdot||$ is the L_1 norm. We assume the following regularity conditions. Some of these conditions are somewhat parallel to those in Carrillo, Chen, and Wu (2010) and Shao (2003, Lemma 5.3) for one-level models, but additional conditions and more complex derivations are required here due to the accommodation of the two-level models with survey weights.

(1) Θ is a compact subset of the Euclidean space R^p.

(2) $\sup_{(i,j,k)} E_\xi\{h^2_{ijk}(Y_{ij}, Y_{ik})\} < \infty$ and $\sup_{1\le i\le N} E_\xi\{||\mathbf{Y}_i||\} < \infty$, where $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iM_i})^{\mathrm{T}}$.

(3) For any given $c > 0$ and a given sequence $\{\mathbf{y}_i\}$ satisfying $||\mathbf{y}_i|| \le c$, the sequence of functions in $\boldsymbol{\theta}$, $\{\boldsymbol{\psi}_{ijk}(\boldsymbol{\theta}; y_{ij}, y_{ik})\}$, is equicontinuous on Θ.

(4) Let $\Delta_T(\boldsymbol{\theta}) = E_\xi E_d\{T^{-1}\mathbf{U}_w(\boldsymbol{\theta})\}$, where $T = \sum_{i=1}^N M_i(M_i - 1)/2$. For any $\epsilon > 0$, there exists $\delta_\epsilon > 0$ such that $\inf_{||\boldsymbol{\theta}-\boldsymbol{\theta}_0||>\epsilon} ||\Delta_T(\boldsymbol{\theta})|| > \delta_\epsilon$.

(5) There exists a $\widehat{\boldsymbol{\theta}}_w \in \Theta$ such that $\mathbf{U}_w(\widehat{\boldsymbol{\theta}}_w) = \mathbf{0}$.

(6) For variable V_{ijk}, write

$$\overline{V} = \frac{1}{T}\sum_{i=1}^{N}\sum_{1\le j<k\le M_i} V_{ijk}.$$

If the V_{ijk} satisfy $\sum_{i=1}^{N}\sum_{1\le j<k\le M_i} V_{ijk}^2/T = O_\xi(1)$, then

$$\frac{1}{T}\sum_{i\in s} w_i \sum_{j<k,j,k\in s(i)} w_{jk|i}V_{ijk} - \overline{V}$$

converges to 0 in design probability as $n \to \infty$.

(7) When the number of clusters in a sample approaches infinity, the number of clusters in the corresponding population tends to infinity as well.

(8) $$\frac{N\ \sup_{i\le N} M_i(M_i-1)}{\sum_{i\le N} M_i(M_i-1)} < \infty \ \text{ as } N \to \infty.$$

Appendix B: Sketched Proof of Theorem 1

Lemma B.1. Under the regularity conditions in Appendix A, we have

$$\sup_{\boldsymbol{\theta}\in\Theta}\Big|\Big|\frac{1}{T}\mathbf{U}_w(\boldsymbol{\theta}) - \Delta_T(\boldsymbol{\theta})\Big|\Big| \xrightarrow{p} 0 \ \text{ as } n \to \infty,$$

where "p" denotes convergence in probability with respect to joint model ξ and sampling design d.

Proof. The detailed derivations are displayed in the web-appendix Section 2.

Proof of Theorem 1. Note that

$$\begin{aligned}\Big|\frac{1}{T}\mathbf{U}_w(\boldsymbol{\theta})\Big| &= \Big|\Delta_T(\boldsymbol{\theta}) + \frac{1}{T}\mathbf{U}_w(\boldsymbol{\theta}) - \Delta_T(\boldsymbol{\theta})\Big| \\ &\ge |\Delta_T(\boldsymbol{\theta})| - \Big|\frac{1}{T}\mathbf{U}_w(\boldsymbol{\theta}) - \Delta_T(\boldsymbol{\theta})\Big|.\end{aligned}$$

By Lemma B.1, for any $\epsilon > 0$, we have

$$\begin{aligned}\inf_{||\boldsymbol{\theta}-\boldsymbol{\theta}_0||>\epsilon}\Big|\frac{1}{T}\mathbf{U}_w(\boldsymbol{\theta})\Big| &\ge \inf_{||\boldsymbol{\theta}-\boldsymbol{\theta}_0||>\epsilon}\Big\{|\Delta_T(\boldsymbol{\theta})| - \Big|\frac{1}{T}\mathbf{U}_w(\boldsymbol{\theta}) - \Delta_T(\boldsymbol{\theta})\Big|\Big\} \\ &\ge \inf_{||\boldsymbol{\theta}-\boldsymbol{\theta}_0||>\epsilon}|\Delta_T(\boldsymbol{\theta})| - \sup_{||\boldsymbol{\theta}-\boldsymbol{\theta}_0||>\epsilon}\Big|\frac{1}{T}\mathbf{U}_w(\boldsymbol{\theta}) - \Delta_T(\boldsymbol{\theta})\Big| \\ &\ge \inf_{||\boldsymbol{\theta}-\boldsymbol{\theta}_0||>\epsilon}|\Delta_T(\boldsymbol{\theta})| - \sup_{\boldsymbol{\theta}\in\Theta}\Big|\frac{1}{T}\mathbf{U}_w(\boldsymbol{\theta}) - \Delta_T(\boldsymbol{\theta})\Big| \\ &\ge \inf_{||\boldsymbol{\theta}-\boldsymbol{\theta}_0||>\epsilon}|\Delta_T(\boldsymbol{\theta})| + o_p(1).\end{aligned}$$

It follows from Assumption 4 that, for any $\epsilon > 0$, there exists $\delta_\epsilon > 0$ such that

$$P_{\xi d}\Big\{\inf_{||\boldsymbol{\theta}-\boldsymbol{\theta}_0||>\epsilon}\Big|\frac{1}{T}\mathbf{U}_w(\boldsymbol{\theta})\Big| > \delta_\epsilon\Big\} \to 1$$

as $n \to \infty$, where the probability $P_{\xi d}$ is evaluated under the model ξ and sampling design d. Noting that $\mathbf{U}_w(\widehat{\boldsymbol{\theta}}_w) = \mathbf{0}$ by Assumption 5, the limit above implies that, for any $\epsilon > 0$, $P_{\xi d}(||\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_w|| \le \epsilon) \to 1$ as $n \to \infty$. This completes the proof that $\widehat{\boldsymbol{\theta}}_w \xrightarrow{p} \boldsymbol{\theta}$.

References

Arellano-Valle, R. B. and Azzalini, A. (2008). The centred parametrization for the multivariate skew-normal distribution. *J. Multivariate Anal.* **99**, 1362-1382.

Asparouhov, T. (2006). Generalized multi-level modeling with sampling weights. *Communications in Statistics - Theory and Methods* **35**, 439-460.

Azzalini, A. and Capitanio, A. (1999). Statistical applications of the multivariate skew normal distribution. *J. Roy. Statist. Soc. Ser. B* **61**, 579-602.

Besag, J. (1974). Spatial interaction and the statistical analysis or lattice systems (with discussion). *J. Roy. Statist. Soc. Ser. B* **36**, 192-236.

Bickel, P. J. and Freedman, D. A. (1984). Asymptotic normality and the bootstrap in stratified sampling. *Ann. Statist.* **12**, 470-482.

Binder, D. A. (1983). On the variance of asymptotically normal estimators form complex surveys. *Int. Statist. Rev.* **51**, 279-292.

Carrillo, I. A., Chen, J. and Wu, C. (2010). The pseudo-GEE approach to the analysis of longitudinal surveys. *Canad. J. Statist.* **38**, 540-554.

Chen, J. and Rao, J. N. K. (2007). Asymptotic normality under two-phase sampling designs. *Statist. Sinica* **17**, 1047-1064.

Cochran, W. G. (1977). *Sampling Techniques.* 3rd Edition. Wiley, New York.

Demnati, A. and Rao, J. N. K. (2010). Linearization variance estimators for model parameters from complex survey data. *Survey Method.* **36**, 193-201.

Grilli, L. and Pratesi, M. (2004). Weighted estimation in multi-level ordinal and binary models in the presence of informative sampling designs. *Survey Method.* **30**, 93-103.

Haziza, D., Mecatti, F. and Rao, J. N. K. (2008). Evaluation of some approximate variance estimators under the Rao-Sampford unequal probability sampling design. *Metron* **66**, 91-108.

He, W. and Yi, G. Y. (2011). A pairwise likelihood method for correlated binary data with/without missing observations under generalized partially linear single-index models. *Statist. Sinica* **21**, 207-229.

Heagerty, P. J. and Lele, S. R. (1998). A composite likelihood approach to binary spatial data. *J. Amer. Statist. Assoc.* **93**, 1099-1111.

Joe, H. and Lee, Y. (2009). On weighting of bivariate margins in pairwise likelihood. *J. Multivariate Anal.* **100**, 670-685.

Korn, E. L. and Graubard, B. I. (2003). Estimating variance components by using survey data. *J. Roy. Statist. Soc. Ser. B* **65**, 175-190.

Kovacevic, M. S. and Rai, S. N. (2003). A pseudo maximum likelihood approach to multi-level modeling of survey data. *Comm. Statist. Theory Methods* **32**, 103-121.

Li, H. and Yi, G. Y. (2013). A pairwise likelihood approach for longitudinal data with missing observations in both response and covariates. *Comput. Statist. Data Anal.* **68**, 66-81.

Lin, T. and Lee, J. (2008). Estimation and prediction in linear mixed models with skew-normal random effects for longitudinal data. *Statist. Medicine* **27**, 1490-1507.

Lindsay, B. G. (1988). Composite likelihood methods. *Contemporary Math.* **80**, 220-239.

Lindsay, B. G., Yi, G. Y. and Sun, J. (2011). Issues and strategies in the selection of composite likelihoods. *Statist. Sinica* **21**, 71-105.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data.* Springer.

Nott, D. and Rydén, T. (1999). Pairwise likelihood methods for inference in image models. *Biometrika* **86**, 661-676.

Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H. and Rasbash, J. (1998). Weighting for unequal selection probabilities in multi-level models. *J. Roy. Statist. Soc. Ser. B* **60**, 23-56.

Pfeffermann, D., Moura, F. and Silva, P. (2006). Multi-level modeling under informative sampling. *Biometrika* **93**, 943-959.

Rabe-Hesketh, S. and Skrondal, A. (2006). Multilevel modelling of complex survey data. *J. Roy. Statist. Soc. Ser. A* **169**, 805-827.

Rao, J. N. K. (1965). On two simple schemes of unequal probability sampling without replacement. *J. Indian Statist. Assoc.* **3**, 173-180.

Rao, J. N. K. and Roberts, G. (1998). Discussion on the papers by Firth and Bennett and Pfeffermann et al. *J. Roy. Statist. Soc. Ser. B* **60**, 50-51.

Rao, J. N. K., Verret, F. and Hidiroglou, M. A. (2013). A weighted composite likelihood approach to inference for two-level models from survey data. *Survey Method.* **39**, 263-282.

Rao, J. N. K., Wu, C. F. J. and Yue, K. (1992). Some recent work on resampling methods for complex surveys. *Survey Method.* **18**, 209-217.

Rust, K. F. and Rao, J. N. K. (1996). Variance estimation for complex surveys using replication techniques. *Statist. Methods in Medical Res.* **5**, 283-310.

Sampford, M. R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika* **54**, 499-513.

Shao, J. (2003). *Mathematical Statistics.* 2nd edition. Springer-Verlag New York, Inc.

Skinner, C. J. (1989). Domain means, regression and multivariate analysis. In *Analysis of Complex Surveys* (Edited by C. J. Skinner, D. Holt and T. M. F. Smith), 59-87. Wiley, New York.

Stapleton, L. (2002). The incorporation of sample weights into multilevel structural equation models. *Structural Eq. Model* **9**, 475-502.

Varin, C. (2008). On composite marginal likelihoods. *Adv. Statist. Anal.* **92**, 1-28.

Varin, C., Reid, N. and Firth, D. (2011). An overview of composite likelihood methods. *Statist. Sinica* **21**, 5-24.

Yi, G. Y., Zeng, L. and Cook, R. J. (2011). A robust pairwise likelihood method for incomplete longitudinal binary data arising in clusters. *Canad. J. Statist.* **39**, 34-51.

Department of Statistics and Actuarial Science, University of Waterloo, 200 University Avenue West, Waterloo, Ontario, Canada N2L 3G1.

E-mail: yyi@uwaterloo.ca

School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada K1S 5B6.

E-mail: jrao@math.carleton.ca

Departments of Oncology and Community Health Sciences, University of Calgary, Calgary, Alberta, Canada T2N 1N4.

E-mail: haocheng.li@ucalgary.ca

(Received December 2013; accepted April 2015)