A Composite Likelihood Approach to Analysis of Survey Data with Sampling Weights Incorporated under Two-Level Models

Grace Y. Yi[13], JNK Rao[2] and Haocheng Li[1]

1. University of Waterloo, Waterloo, Canada
2. Carleton University, Ottawa, Canada
3. Corresponding author: Grace Y. Yi, *E-mail: yyi@uwaterloo.ca*

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. On the other hand, available methods that are valid for genera 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 reported have demonstrated that the method performs well in estimating the model parameters. Moreover, this research has important implication: It provides a particular scenario to showcase the unique merit of the composite likelihood method for which the likelihood method would not work.

Key words: Composite likelihood; Complex sampling design; Design-based inference; Multi-level model; Super-population model; Variance estimation.

1 Introduction

Multi-level models provide a flexible framework to include auxiliary variables related to survey design features. When carrying out inference about the model parameters, it is important to accommodate sampling characteristics, such as stratification, clustering, and unequal selection probabilities; otherwise, misleading or erroneous results may result. 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; Kovacevic and Rai, 2003; Grilli and Pratesi, 2004; Pfeffermann et al., 2006; Asparouhov, 2006; Rabe-Hesketh and Skrondal, 2006; Rao et al., 2010), research in this area remains relatively unexplored.

In this paper, we address this important problem by exploring a unified inferential procedure for multi-level models featuring survey data with sampling probabilities incorporated. Our approach is based on the composite likelihood formulation (Lindsay, 1988; Lindsay et al., 2011). Rao et al. (2010) 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 of the Rao et al. (2010) method.

2 Notation and Framework

We consider the case of 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 subject j in cluster i, and $\mathbf{x}_{ij}$ be the associated covariate vector, $i = 1, ..., N$, and $j = 1, ..., M_i$. Correspondingly, the super-population model from which this finite population is generated is assumed to match the design two-level structure. We assume that given cluster i and random effects $\mathbf{u}_i$, Y_{ij} are assumed to be independently distributed as

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

where $f_{y|u}$ is a known density function and $\boldsymbol{\theta}_y$ is the associated parameter vector. In the second step we further model random effects by assuming that

$$\mathbf{u}_i \text{ has a density function } f_u \tag{2}$$

where f_u often has a given parametric form that is indexed by the parameter $\boldsymbol{\theta}_u$. This model formulation covers both linear two-level models (i.e., linear mixed models) and generalized linear two-level models (generalized linear mixed models). With informative sampling of clusters and of elements within sampled clusters, the population model above may not hold for the sample. In that case, standard methods for multi-level models that ignore the design and assume model (1) with (2) holds 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 the next section, we tackle this problem using the weighted composite likelihood framework in order to attain both validity and robustness of results.

3 Estimation based on Composite Likelihood

Let $L_{ij} = f(y_{ij}|\mathbf{x}_{ij})$ be the density 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 density for paired responses (Y_{ij}, Y_{ik}), and this is determined by

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

Let $\ell_{ij} = \log L_{ij}$, and $\ell_{ijk} = \log L_{ijk}$.

A "census" composite likelihood can be formulated based on the marginal pairwise distributions (Lindsay et al., 2011):

$$C(\boldsymbol{\theta}) = \prod_{i=1}^{N} \prod_{j<k} L_{ijk}.$$

A "census" log pairwise likelihood under the assumed two-level model given by (1) and (2) is obtained as

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

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

$$\ell_{wc}(\boldsymbol{\theta}) = \sum_{i \in s} w_i \sum_{j<k, j,k \in s(i)} w_{jk|i} \ell_{ijk}, \tag{3}$$

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

$$\mathbf{U}_{wc}(\boldsymbol{\theta}) = \frac{\partial \ell_{wc}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i \in s} w_i \mathbf{U}_{iwc}(\boldsymbol{\theta}) = \mathbf{0} \tag{4}$$

for $\boldsymbol{\theta}$ leads to the weighted composite likelihood estimator, $\widehat{\boldsymbol{\theta}}_w$, of $\boldsymbol{\theta}$, where $\mathbf{U}_{iwc}(\boldsymbol{\theta}) = \sum_{j<k, j,k \in s(i)} w_{jk|i} \mathbf{s}_{ijk}$, and $\mathbf{s}_{ijk} = \partial \ell_{ijk} / \partial \boldsymbol{\theta}$.

One notices from (3) and (4) that to implement the proposed method, we need the within-cluster joint inclusion probabilities $\pi_{jk|i} = w_{jk|i}^{-1}$, in addition to the inclusion probabilities $\pi_i = w_i^{-1}$. This information is often available in many settings, such as simple random or stratified random sampling within clusters, or when the within cluster sampling fraction is small. If such information is not available, one may employ an approximation to $\pi_{jk|i}$. When sampling within clusters 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 et al. (2008) for details.

Now we explain that $E_\xi E_d\{U_{wc}(\boldsymbol{\theta})\} = 0$. By the nature of the two-stage design weights $w_{j|i}$ and w_i, it is seen that the inner expectation $E_d\{U_{wc}(\boldsymbol{\theta})\}$ recovers the census composite score function, $U_c(\boldsymbol{\theta}) = (\partial/\partial\boldsymbol{\theta})\{\ell_c(\boldsymbol{\theta})\}$. 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 both the design and model perspectives. 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-cluster sizes, m_i, are small.

4 Variance Estimation

The covariance matrix of the estimator $\widehat{\boldsymbol{\theta}}_w$ is given by

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

Let $\boldsymbol{\theta}_U = E_d(\widehat{\boldsymbol{\theta}}_w)$, then $\boldsymbol{\theta}_U$ can be viewed as a finite population (or census) quantity which is unbiasedly estimated by the estimator $\widehat{\boldsymbol{\theta}}_w$. As discussed by Demnati and Rao (2010) and Carrillo et al. (2010), if $\mathrm{cov}_\xi(\boldsymbol{\theta}_U)$ has the order of $1/N$ and the sampling fraction n/N is small, then we can approximate $\mathrm{cov}_{\xi d}(\widehat{\boldsymbol{\theta}}_w)$ by the term $E_\xi\{\mathrm{cov}_d(\widehat{\boldsymbol{\theta}}_w)\}$. That is,

$$\mathrm{cov}_{\xi d}(\widehat{\boldsymbol{\theta}}_w) \approx E_\xi\{\mathrm{cov}_d(\widehat{\boldsymbol{\theta}}_w)\}, \tag{5}$$

and this 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 $\mathrm{cov}_{\xi d}(\widehat{\boldsymbol{\theta}}_w)$.

The design-based covariance $\mathrm{cov}_d(\widehat{\boldsymbol{\theta}}_w)$ can be estimated using the Taylor series expansion that is similar to Binder (1983). That is, we use the approximation:

$$\mathrm{cov}_d(\widehat{\boldsymbol{\theta}}_w) \approx \{\boldsymbol{\Gamma}_c(\boldsymbol{\theta}_N)\}^{-1}\mathrm{cov}_d\{\mathbf{U}_{wc}(\boldsymbol{\theta}_N)\}\{\boldsymbol{\Gamma}_c(\boldsymbol{\theta}_N)\}^{-1\mathrm{T}}. \tag{6}$$

A precise evaluation of (6) is generally difficult as it requires fourth order within-cluster inclusion probabilities. We therefore follow the customary practice of treating the sample clusters 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 can 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 $\mathbf{U}_{iwc}(\boldsymbol{\theta}) = \sum_{j<k,j,k\in s(i)} w_{jk|i}\mathbf{s}_{ijk}(\boldsymbol{\theta})$. Consequently, we estimate the covariance matrix of $\mathbf{U}_{wc}(\boldsymbol{\theta}_N)$ as

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

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

$$\widehat{\mathrm{cov}}_d\{\mathbf{U}_{wc}(\boldsymbol{\theta}_N)\} = \frac{n}{(n-1)}\sum_{i\in s} w_i^2\mathbf{U}_{iwc}\mathbf{U}_{iwc}^{\mathrm{T}} \tag{7}$$

where $w_i = 1/\pi_i$ and $\mathbf{U}_{iwc} = \mathbf{U}_{iwc}(\widehat{\boldsymbol{\theta}}_w)$. We will use (7) as an approximate estimator of the covariance matrix $\mathrm{cov}_{\xi d}(\widehat{\boldsymbol{\theta}}_w)$. This approximation should perform well if the level 2 sampling fraction n/N is small; otherwise it will lead to overestimation bias.

It now follows from (5), (6) and (7) that an approximate estimator of $\mathrm{cov}_{\xi d}(\widehat{\boldsymbol{\theta}}_w)$ is given by

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

Acknowledgement

This research was supported by the Natural Sciences and Engineering Research Council of Canada.

References

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

Binder, D. A. (1983). On the variance of asymptotically normal estimators form complex surveys. *International Statistical Review*, 51:279–292.

Carrillo, I. A., Chen, J., and Wu, C. (2010). The pseudo-gee approach to the analysis of longitudinal surveys. *The Canadian Journal of Statistics*, 38:540–554.

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

Kovacevic, M. S. and Rai, S. N. (2003). A pseudo maximum likelihood approach to multilevel modeling of survey data. *Communications in Statistics - Theory and Methods*, 32:103–121.

Lindsay, B. G. (1988). Composite likelihood methods. *Contemporary Mathematics*, 80:220–239.

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

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

Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., and Rasbash, J. (1998). Weighting for unequal selection probabilities in multi-level models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60:23–56.

Rabe-Hesketh, S. and Skrondal, A. (2006). Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169:805–827.

Rao, J. N. K. (1965). On two simple schemes of unequal probability sampling without replacement. *Journal of the Indian Statistical Association*, 3:173–180.

Rao, J. N. K., Verret, F., and Hidiroglou, M. A. (2010). A weighted estimating equations approach to inference for two-level models from survey data. *Proceedings of the Survey Methods Section, SSC Annual Meeting, May 2010*.

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

Skinner, C. J. (1989). *Domain means, regression and multivariate analysis in Analysis of Complex Surveys*. New York: Wiley.

Stapleton, L. (2002). The incorporation of sample weights into multilevel structural equation models. *Structural Equation Modeling*, 9:475–502.