

## Article

# A weighted composite likelihood approach to inference for two-level models from survey data

by J.N.K. Rao, François Verret and Mike A. Hidioglou

January 2014



## How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, [www.statcan.gc.ca](http://www.statcan.gc.ca).

You can also contact us by

**email** at [infostats@statcan.gc.ca](mailto:infostats@statcan.gc.ca),

**telephone**, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following toll-free numbers:

- |   |                |
|---|----------------|
| • Statistical Information Service                             | 1-800-263-1136 |
| • National telecommunications device for the hearing impaired | 1-800-363-7629 |
| • Fax line  | 1-877-287-4369 |

## Depository Services Program

- |                  |                |
|------------------|----------------|
| • Inquiries line | 1-800-635-7943 |
| • Fax line       | 1-800-565-7757 |

## To access this product

This product, Catalogue no. 12-001-X, is available free in electronic format. To obtain a single issue, visit our website, [www.statcan.gc.ca](http://www.statcan.gc.ca), and browse by “Key resource” > “Publications.”

## Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on [www.statcan.gc.ca](http://www.statcan.gc.ca) under “About us” > “The agency” > “Providing services to Canadians.”

Published by authority of the Minister responsible for  
Statistics Canada

© Minister of Industry, 2014.

All rights reserved. Use of this publication is governed by the  
Statistics Canada Open Licence Agreement ([http://www.  
statcan.gc.ca/reference/licence-eng.html](http://www.statcan.gc.ca/reference/licence-eng.html)).

Cette publication est aussi disponible en français.

## Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

## Standard symbols

The following symbols are used in Statistics Canada publications:

- |                |  |
|----------------|--|
| .              | not available for any reference period   |
| ..             | not available for a specific reference period  |
| ...            | not applicable   |
| 0              | true zero or a value rounded to zero   |
| 0 <sup>s</sup> | value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded |
| P              | preliminary  |
| r              | revised  |
| X              | suppressed to meet the confidentiality requirements of the <i>Statistics Act</i>                                   |
| E              | use with caution   |
| F              | too unreliable to be published   |
| *              | significantly different from reference category ( $p < 0.05$ )   |

# A weighted composite likelihood approach to inference for two-level models from survey data

J.N.K. Rao, François Verret and Mike A. Hidiroglou<sup>1</sup>

## Abstract

Multi-level models are extensively used for analyzing survey data with the design hierarchy matching the model hierarchy. We propose a unified approach, based on a design-weighted log composite likelihood, for two-level models that leads to design-model consistent estimators of the model parameters even when the within cluster sample sizes are small provided the number of sample clusters is large. This method can handle both linear and generalized linear two-level models and it requires level 2 and level 1 inclusion probabilities and level 1 joint inclusion probabilities, where level 2 represents a cluster and level 1 an element within a cluster. Results of a simulation study demonstrating superior performance of the proposed method relative to existing methods under informative sampling are also reported.

**Key Words:** Composite likelihood; Inclusion probabilities; Informative sampling; Multi-level models.

## 1 Introduction

Data collected from large-scale socio-economic, health and other surveys are extensively used for analysis purposes, such as inference on the regression parameters of linear and logistic linear regression population models. Ignoring the survey design features (such as stratification, clustering and unequal selection probabilities) can lead to erroneous inferences on model parameters because of sample selection bias caused by informative sampling. It is tempting to expand the models by including among the auxiliary variables all the design variables that define the selection process at the various levels and then ignore the design and apply standard methods to the expanded model. The main difficulties with this approach are the following (Pfeffermann and Sverchkov 2003): (1) Not all design variables may be known or accessible to the analyst; (2) Too many design variables can lead to difficulties in making inference from the expanded model; (3) The expanded model may no longer be of scientific interest to the analyst. On the other hand, the design-based approach can provide asymptotically valid repeated sampling inferences without changing the analyst's model. A unified approach based on the survey weighted estimating equations leads to design-consistent estimators of the "census" or finite population parameters which in turn estimate the associated model parameters. Further, re-sampling methods, such as the jackknife and the bootstrap for survey data, can provide valid variance estimators and associated inferences on the census parameters. The same methods may also be applicable to inference on the model parameters, in many cases of large-scale surveys. In other cases, it is necessary to estimate the model variance of the census parameters from the sample. The estimator of the total variance is then given by the sum of this estimator and the re-sampling variance estimator. Beaumont and Charest (2010) extended the bootstrap to estimate the total variance associated with the model parameters. We refer the reader to Rao *et al.* (2010) for an overview of methods for making inference on regression parameters from complex survey data.

---

1. J.N.K. Rao, School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada, K1S 5B6. E-mail: jrao@math.carleton.ca; François Verret, Statistics Canada, 15 B, R.-H.-Coats Building, Ottawa, Ontario, Canada, K1A 0T6. E-mail: francois.verret@statcan.gc.ca; Mike A. Hidiroglou, Statistics Canada, 16 D, R.-H.-Coats Building, Ottawa, Ontario, Canada, K1A 0T6. E-mail: mike.hidiroglou@statcan.gc.ca.

In this paper, our focus is on making design-based inference on the variance component parameters and regression parameters of multi-level models from data obtained from multi-stage sampling designs corresponding to the levels of the model. For example, in an education study of students, schools (first-stage sampling units) may be selected with probabilities proportional to school size and students (second-stage units) within selected schools by stratified random sampling. Again, ignoring the survey design and using traditional methods for multi-level models can lead to erroneous inferences in the presence of sample selection bias. In the design-based approach, estimation of variance component parameters of the model is more difficult than that of regression parameters. Past work on multi-level models for survey data is summarized in Section 2. Our main purpose is to present a unified approach to making inference for general multi-level models from survey data, based on a weighted log composite likelihood approach (Section 4). The proposed methods lead to asymptotically valid inferences on the variance component parameters even when the within-cluster sample sizes are small, provided the number of sample clusters is large, unlike some of the existing methods summarized in Section 2. Limited simulation results are presented in Section 5.

## 2 Two-level models: Past work

### 2.1 Two-level models

Multi-level (or hierarchical) models are extensively used in social sciences, education, health and other areas to analyze survey data with a hierarchical structure. Here we focus on two-level models associated with two-stage sampling of clusters (level 2): a sample,  $s$ , of level 2 units,  $i$ , is selected according to a specified design and then a sample,  $s(i)$ , of elements (or level 1 units),  $j$ , is selected from each sampled level 2 unit  $i$  according to another specified design. We assume, following the literature on multi-level models for survey data, that the model matches the design hierarchy, as in the example of an educational survey of students. However, in some multipurpose surveys, the design hierarchical structure could be quite different from the model hierarchy. For example, the Canadian National Longitudinal Survey of Children and Youth uses a multi-stage design where the stages are geographical areas, households within an area and students within a household, whereas an educational multilevel model may include as levels students, classes, schools and school boards (Rao and Roberts 1998). Since the design clusters cut across the model clusters for such surveys, it is difficult to develop a suitable design-weighted method of inference on the model parameters that can handle informative sampling of clusters and or elements within sampled clusters. Under informative sampling, the assumed model for the population may not hold for the sample.

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  $i$ . A two-level super-population model is given by

$$y_{ij} \mid \mathbf{x}_{ij}, \mathbf{v}_i \sim_{ind} f(y_{ij} \mid \mathbf{x}_{ij}, \mathbf{v}_i, \boldsymbol{\theta}_1), \quad \mathbf{v}_i \sim_{iid} f(\mathbf{v}_i \mid \boldsymbol{\theta}_2), \quad i = 1, \dots, N; j = 1, \dots, M_i, \quad (2.1)$$

where  $y_{ij}$  and  $\mathbf{x}_{ij} = (x_{ij0}, \dots, x_{ij,p-1})^T$  are the response and a  $p$ -vector of covariate values associated with element  $j$  within cluster  $i$  and  $x_{ij0} = 1$ ,  $\mathbf{v}_i$  denotes a level 2 random effect, and  $\boldsymbol{\theta}_1$  and  $\boldsymbol{\theta}_2$  denote the

parameters associated with the two stages of the assumed model. Here  $f(y_{ij} | \mathbf{x}_{ij}, \mathbf{v}_i, \boldsymbol{\theta}_1)$  and  $f(\mathbf{v}_i | \boldsymbol{\theta}_2)$  are specified density functions of  $y_{ij}$  given  $\mathbf{x}_{ij}$  and  $\mathbf{v}_i$  and of  $\mathbf{v}_i$ , respectively. Note that in model (2.1), the responses  $y_{ij}$  for a given  $i$  are assumed to be conditionally independent given the random effect  $\mathbf{v}_i$  but they are correlated marginally due to the common  $\mathbf{v}_i$ . The model formulation (2.1) covers both linear two-level models and generalized linear two-level models. Under informative sampling of clusters and/or elements within sampled clusters, standard methods for multi-level models that ignore the design and assume that model (2.1) holds for the sample can lead to asymptotically biased estimators of model parameters  $\boldsymbol{\theta}_1$  and  $\boldsymbol{\theta}_2$  (Pfeffermann *et al.* 1998).

### Special cases

(1) A simple nested error mean model that is often used in simulation studies related to two-level models is given by

$$y_{ij} = \mu + v_i + e_{ij}, e_{ij} \sim_{iid} N(0, \sigma_e^2), v_i \sim_{iid} N(0, \sigma_v^2), \quad (2.2)$$

where  $i = 1, \dots, N; j = 1, \dots, M_i$ . Model (2.2) may be written in the form (2.1) as

$$y_{ij} | v_i \sim_{ind} N(\mu + v_i, \sigma_e^2), v_i \sim_{iid} N(0, \sigma_v^2), \boldsymbol{\theta}_1 = (\mu, \sigma_e^2), \boldsymbol{\theta}_2 = \sigma_v^2.$$

Marginally,  $y_{ij} \sim N(\mu, \sigma_v^2 + \sigma_e^2)$  but  $y_{ij}$  and  $y_{ik}$  ( $j \neq k$ ) are correlated:  $\text{corr}(y_{ij}, y_{ik}) = \rho = \sigma_v^2 / (\sigma_v^2 + \sigma_e^2), j \neq k$ .

(2) A linear two-level model, often used in practice, is given by

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta}_i + e_{ij}, i = 1, \dots, N; j = 1, \dots, M_i, \quad (2.3)$$

where  $\boldsymbol{\beta}_i = \boldsymbol{\beta} + \mathbf{v}_i, \mathbf{v}_i \sim_{iid} N_p(\mathbf{0}, \boldsymbol{\Sigma}_v), i = 1, \dots, N$  and  $e_{ij} \sim_{iid} N(0, \sigma_e^2)$ . This model may also be expressed in the form (2.1) as

$$y_{ij} | \mathbf{x}_{ij}, \mathbf{v}_i \sim_{ind} N(\mathbf{x}_{ij}^T \boldsymbol{\beta} + \mathbf{x}_{ij}^T \mathbf{v}_i, \sigma_e^2), \mathbf{v}_i \sim_{iid} N_p(\mathbf{0}, \boldsymbol{\Sigma}_v) \quad (2.4)$$

where  $\boldsymbol{\theta}_1 = (\boldsymbol{\beta}^T, \sigma_e^2)^T$  and  $\boldsymbol{\theta}_2$  is the vector of  $p(p+1)/2$  distinct elements of  $\boldsymbol{\Sigma}_v$ . Marginally,  $y_{ij} \sim N(\mathbf{x}_{ij}^T \boldsymbol{\beta}, \mathbf{x}_{ij}^T \boldsymbol{\Sigma}_v \mathbf{x}_{ij} + \sigma_e^2)$ , but  $y_{ij}$  and  $y_{ik}$  ( $j \neq k$ ) are correlated through the common random effect  $\mathbf{v}_i$ . However, in the case of a generalized linear two-level model, the marginal distribution of  $y_{ij}$  generally does not yield a closed-form expression; for example, in the case of a logistic linear two-level model for binary responses.

## 2.2 Point estimation

The “census” or population log-likelihood under the assumed two-level model (2.1) is given by

$$\log L(\boldsymbol{\theta}) = \sum_{i=1}^N \log L_i(\boldsymbol{\theta}) \equiv \sum_{i=1}^N l_i(\boldsymbol{\theta}) = l(\boldsymbol{\theta}), \quad (2.5)$$

where  $\boldsymbol{\theta}$  is the vector with elements  $\boldsymbol{\theta}_1$  and  $\boldsymbol{\theta}_2$ , and

$$L_i(\boldsymbol{\theta}) = \int \exp \left[ \sum_{j=1}^{M_i} \log f(y_{ij} | \mathbf{x}_{ij}, \mathbf{v}_i, \boldsymbol{\theta}_1) \right] f(\mathbf{v}_i | \boldsymbol{\theta}_2) d\mathbf{v}_i \quad (2.6)$$

see Asparouhov (2006) and Rabe-Hesketh and Skrondal (2006). The census score function  $\mathbf{U}(\boldsymbol{\theta}) = \partial l(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}$  satisfies  $E_m \{\mathbf{U}(\boldsymbol{\theta})\} = \mathbf{0}$ , where  $E_m$  denotes the model expectation. The census parameter  $\boldsymbol{\theta}_N$  is defined as the unique solution to  $\mathbf{U}(\boldsymbol{\theta}) = \mathbf{0}$  and  $\boldsymbol{\theta}_N$  is model consistent for  $\boldsymbol{\theta}$ , where  $\boldsymbol{\theta}_N$  is the vector with elements  $\boldsymbol{\theta}_{1N}$  and  $\boldsymbol{\theta}_{2N}$ .

Let the sample consist of  $n$  clusters with  $m_i$  elements from sample cluster  $i$ . Let  $\pi_i$  and  $\pi_{ji}$  respectively denote the level 2 and level 1 inclusion probabilities associated with cluster  $i$  and element  $j$  within cluster  $i$ . Then the level 2 and level 1 weights are given by  $w_i = \pi_i^{-1}$  and  $w_{ji} = \pi_{ji}^{-1}$  respectively. Asparouhov (2006) and Rabe-Hesketh and Skrondal (2006) proposed a weighted sample pseudo log-likelihood obtained by replacing  $\sum_{j=1}^{M_i} (\cdot)$  in (2.6) by  $\sum_{j \in s_i} w_{ji} (\cdot)$  and  $\sum_{i=1}^N (\cdot)$  in (2.5) by  $\sum_{i \in s} w_i (\cdot)$ , where  $s$  denotes the sample of clusters and  $s(i)$  denotes the sample of elements within clusters  $i \in s$ . It is given by

$$\tilde{l}_w(\boldsymbol{\theta}) = \sum_{i \in s} w_i \tilde{l}_{wi}(\boldsymbol{\theta}) \quad (2.7)$$

where  $\tilde{l}_{wi}(\boldsymbol{\theta}) = \log \tilde{L}_{wi}(\boldsymbol{\theta})$  and

$$\tilde{L}_{wi}(\boldsymbol{\theta}) = \int \exp \left[ \sum_{j \in s(i)} w_{ji} \log f(y_{ij} | \mathbf{x}_{ij}, \mathbf{v}_i, \boldsymbol{\theta}_1) \right] f(\mathbf{v}_i | \boldsymbol{\theta}_2) d\mathbf{v}_i. \quad (2.8)$$

Maximizing the pseudo log-likelihood  $\tilde{l}_w(\boldsymbol{\theta})$ , given by (2.7), we get a pseudo maximum likelihood (PML) estimator  $\tilde{\boldsymbol{\theta}}_w$ . Computational details are discussed in Asparouhov (2006) and Rabe-Hesketh and Skrondal (2006). In the special case of linear two-level models, Pfeiffermann *et al.* (1998) used an iterative generalized least squares method proposed by Goldstein (1986). Note that we need both level 1 and level 2 weights to compute  $\tilde{\boldsymbol{\theta}}_w$ , unlike in the case of marginal models that require only the unconditional element weights  $w_{ij} = w_i w_{ji}$ .

Design consistency of the PML estimator  $\tilde{\boldsymbol{\theta}}_{2w}$  of the census parameter  $\boldsymbol{\theta}_{2N}$  or design-model consistency of  $\tilde{\boldsymbol{\theta}}_{2w}$  as an estimator of the model parameter  $\boldsymbol{\theta}_2$  requires that both the number of sample clusters,  $n$ , and the within cluster sample sizes,  $m_i$ , tend to infinity, even in the linear case. Also, the relative bias of the estimators will be considerable when  $m_i$  are small. To remedy this problem, several weight-scaling methods have been proposed in the literature. In particular, level 1 weights  $w_{ji}$  in (2.8) are scaled by a factor  $k_{1i}$  before maximizing the pseudo log-likelihood (2.7). We consider only two weight-scaling methods here, denoted A and A1 (Asparouhov 2006). Method A uses

$$k_{1i} = m_i / \sum_{j \in s(i)} w_{ji} \tag{2.9}$$

In method A1,  $k_{1i}$  is the same as in method A but level 2 weights  $w_i$  are also scaled by the factor  $k_{2i} = 1/k_{1i}$  to offset level 1 weight scaling. Asparouhov (2006) mentioned the use of accelerated EM algorithm for calculating the PML estimator  $\tilde{\theta}_w$  with M plus 3: www.Statmodel.com: Muthén and Muthén, 1998-2005.

### 2.3 Variance estimation

Turning to variance estimation, Asparouhov (2006) proposed a Taylor linearization “sandwich” variance estimator of  $\tilde{\theta}_w$ . It is given by

$$v_L(\tilde{\theta}_w) = (\tilde{\mathbf{I}}_w'')^{-1} \left[ \sum_{i \in s} (k_{2i} w_i)^2 \tilde{\mathbf{I}}_{wi}' (\tilde{\mathbf{I}}_{wi}')^T \right] (\tilde{\mathbf{I}}_w')^{-1}, \tag{2.10}$$

where  $\tilde{\mathbf{I}}_w'$  and  $\tilde{\mathbf{I}}_w''$  respectively denote the first derivative vector and the second derivative matrix of  $\tilde{l}_w(\theta)$  evaluated at  $\theta = \tilde{\theta}_w$ , and  $\tilde{\mathbf{I}}_{wi}'$  is the first derivative of  $\tilde{l}_{wi}(\theta)$  evaluated at  $\theta = \tilde{\theta}_w$ . If the level 2 sampling fraction is small, then  $v_L(\tilde{\theta}_w)$  tracks the variance of  $\tilde{\theta}_w$  well, but not the MSE of  $\tilde{\theta}_w$  if the relative bias of  $\tilde{\theta}_w$  is large.

Kovacevic *et al.* (2006) studied bootstrap variance estimators for  $\tilde{\theta}_w$ . They considered two options: options 1 and 2. In option 1, level 2 bootstrap weights  $w_i(b)$ , based on the Rao, Wu and Yue (1992) method, are used and level 1 weights are not changed, *i.e.*,  $w_{ji}(b) = w_{ji}$ , where  $b = 1, \dots, B$  denote the  $B$  bootstrap samples. For option 2, the Rao, Wu and Yue (1992) bootstrap method is applied to both level 1 and level 2, and the level 1 bootstrap weights are rescaled. Replacing the weights  $w_i$  and  $w_{ji}$  by  $w_i(b)$  and  $w_{ji}(b)$  in (2.7) and (2.8), bootstrap PML estimators  $\tilde{\theta}_w(b)$ ,  $b = 1, \dots, B$  are obtained and the resulting bootstrap variance estimator is given by

$$v_{Boot}(\tilde{\theta}_w) = \frac{1}{B} \sum_{b=1}^B [\tilde{\theta}_w(b) - \tilde{\theta}_w][\tilde{\theta}_w(b) - \tilde{\theta}_w]^T. \tag{2.11}$$

A simulation study of (2.11), based on the simple mean model (2.2), showed that option 1 may lead to underestimation of the variance of  $\tilde{\sigma}_{ew}^2$ . Option 2 performed better than option 1. Grilli and Pratesi (2004) studied an alternative bootstrap method for variance estimation.

## 3 Design-weighted estimating equations

In Sections 3 and 4 we study methods of generating design-weighted estimating equations for the model parameters of multi-level models that lead to design-model consistent estimators, even in the case of small within-cluster sample sizes. The proposed methods depend only on the first order inclusion probabilities  $\pi_i$  and  $\pi_{ji}$  and the joint inclusion probabilities  $\pi_{jki}$  within clusters. Section 3 introduces a

simple moment-based weighted estimating equations approach applicable to linear nested error regression models. A unified method, based on weighted log composite likelihoods, is proposed in Section 4. This method can handle linear and generalized linear multi-level methods, unlike the moment-based method, and it leads to design-model consistent estimators. It also depends only on  $\pi_i$ ,  $\pi_{j|i}$  and  $\pi_{jk|i}$ .

### 3.1 Point estimation

We first illustrate the weighted estimating equations approach, using the simple mean model (2.2). Here our interest is to estimate  $\boldsymbol{\theta} = (\mu, \sigma_v^2, \sigma_e^2)^T$  from a two-stage cluster sampling design matching the model hierarchy. We have chosen the following three estimating functions (EF) for this purpose:

$$u_1(y_{ij}, \boldsymbol{\theta}) = y_{ij} - \mu, \quad (3.1)$$

$$u_2(y_{ij}, \boldsymbol{\theta}) = (y_{ij} - \mu)^2 - (\sigma_v^2 + \sigma_e^2) \quad (3.2)$$

$$u_3(y_{ij}, y_{ik}, \boldsymbol{\theta}) = [(y_{ij} - \mu) - (y_{ik} - \mu)]^2 - 2\sigma_e^2 = z_{ijk}^2 - 2\sigma_e^2, j \neq k, \quad (3.3)$$

where  $z_{ijk} = y_{ij} - y_{ik}$ . The corresponding census estimating equations are given by

$$U_1(\boldsymbol{\theta}) = \sum_{i=1}^N \sum_{j=1}^{M_i} u_1(y_{ij}, \boldsymbol{\theta}) = 0, U_2(\boldsymbol{\theta}) = \sum_{i=1}^N \sum_{j=1}^{M_i} u_2(y_{ij}, \boldsymbol{\theta}) = 0 \quad (3.4)$$

$$U_3(\boldsymbol{\theta}) = \sum_{i=1}^N \sum_{j < k=1}^{M_i} u_3(y_{ij}, y_{ik}, \boldsymbol{\theta}) = 0. \quad (3.5)$$

The resulting census parameter,  $\tilde{\boldsymbol{\theta}}_N$ , is model-consistent for  $\boldsymbol{\theta}$  because the model expectations of the three estimating functions (3.1) – (3.3) are zero. It follows from (3.4) and (3.5) that the design-weighted estimating equations (WEE) are given by

$$\hat{U}_{w1}(\boldsymbol{\theta}) = \sum_{i \in s} w_i \sum_{j \in s(i)} w_{j|i} u_1(y_{ij}, \boldsymbol{\theta}) \equiv \sum_{i \in s} w_i \hat{U}_{w1i}(\boldsymbol{\theta}) = 0 \quad (3.6)$$

$$\hat{U}_{w2}(\boldsymbol{\theta}) = \sum_{i \in s} w_i \sum_{j \in s(i)} w_{j|i} u_2(y_{ij}, \boldsymbol{\theta}) \equiv \sum_{i \in s} w_i \hat{U}_{w2i}(\boldsymbol{\theta}) = 0 \quad (3.7)$$

$$\hat{U}_{w3}(\boldsymbol{\theta}) = \sum_{i \in s} w_i \sum_{j < k \in s(i)} w_{jk|i} u_3(y_{ij}, y_{ik}, \boldsymbol{\theta}) \equiv \sum_{i \in s} w_i \hat{U}_{w3i}(\boldsymbol{\theta}) = 0, \quad (3.8)$$

where  $w_{jk|i} = \pi_{jk|i}^{-1}$ . The WEE estimator,  $\hat{\boldsymbol{\theta}}_w$ , is obtained by solving (3.6) – (3.8). For the mean model, we obtain explicit solutions to WEE as

$$\hat{\mu}_w = \left( \sum_{i \in s} \sum_{j \in s(i)} w_{ij} y_{ij} \right) / \sum_{i \in s} \sum_{j \in s(i)} w_{ij} \equiv \bar{y}_w \quad (3.9)$$

$$\hat{\sigma}_{vw}^2 = \sum_{i \in s} \sum_{j \in s(i)} w_{ij} (y_{ij} - \bar{y}_w)^2 / \sum_{i \in s} \sum_{j \in s(i)} w_{ij} - \hat{\sigma}_{ew}^2 \tag{3.10}$$

$$\hat{\sigma}_{ew}^2 = \sum_{i \in s} w_i \sum_{j < k \in s(i)} w_{jk|i} z_{ijk}^2 / \left( 2 \sum_{i \in s} w_i \sum_{j < k \in s(i)} w_{jk|i} \right), \tag{3.11}$$

where  $w_{ij} = w_i w_{ji}$ . Note that the above moment method is distribution free.

We note that  $\hat{U}_{wt}(\boldsymbol{\theta}), t = 1, 2, 3$  are estimating functions with zero expectation with respect to the design and the model, *i.e.*,  $E_m E_p \{ \hat{U}_{wt}(\boldsymbol{\theta}) \} = 0$ . Using this result, it can be shown that the WEE estimator  $\hat{\boldsymbol{\theta}}_w = (\hat{\boldsymbol{\mu}}_w, \hat{\sigma}_{vw}^2, \hat{\sigma}_{ew}^2)^T$  is design-model consistent for  $\boldsymbol{\theta}$  as the number of level 2 units in the sample,  $n$ , increases, even when the within cluster sample sizes,  $m_i$ , are small. This property does not necessarily hold for the estimators presented in Section 2. The proposed method, however, requires the within-cluster joint inclusion probabilities  $\pi_{jk|i}$ . The latter are readily available for simple random or stratified random sampling within clusters, or when the within cluster sampling fraction is small. Also several good approximations to  $\pi_{jk|i}$  when sampling within clusters is based on unequal probability sampling are also available, and those approximations depend only on the marginal inclusion probabilities  $\pi_{ji}$  (Haziza, Mecatti and Rao 2008). The WEE estimator  $\hat{\boldsymbol{\theta}}_w$  is also design-consistent for  $\tilde{\boldsymbol{\theta}}_N$ , noting that  $E_p \{ \hat{U}_{wt}(\tilde{\boldsymbol{\theta}}_N) \} = 0, t = 1, 2, 3$ .

The choice of estimating functions (3.1) – (3.3) is not necessarily unique. For example, we could replace the previous  $u_2(y_{ij}, \boldsymbol{\theta})$  by  $\tilde{u}_2(y_{ij}, y_{ik}, \boldsymbol{\theta}) = (y_{ij} - \mu)(y_{ik} - \mu) - \sigma_v^2$  in (3.7) and retain (3.6) and (3.8). The resulting WEE estimator is also design-model consistent for  $\boldsymbol{\theta}$  as the number of level 2 units increases. The weighted pairwise composite likelihood approach of Section 4 provides a unified method of generating the estimating functions.

Korn and Graubard (2003) used an alternative approach for the mean model which has some similarities with the proposed approach. Under this approach, “census parameters”,  $S_e^2$  and  $S_v^2$  are first obtained by assuming that the model holds for the finite population. Survey weighted estimators  $\hat{S}_{ew}^2$  and  $\hat{S}_{vw}^2$  of the census parameters are then obtained, assuming  $M_i$  is known for the sampled clusters. The estimator  $\hat{S}_{ew}^2$  is given by

$$\hat{S}_{ew}^2 = \left\{ \frac{1}{2} \sum_{i \in s} (M_i - 1) w_i \left[ \sum_{j < k \in s(i)} w_{jk|i} (y_{ij} - y_{ik})^2 / \sum_{j < k \in s(i)} w_{jk|i} \right] \right\} \left[ \sum_{i \in s} (M_i - 1) w_i \right]^{-1}, \tag{3.12}$$

assuming  $m_i > 1$  for all sampled clusters. Note that (3.12) requires the joint inclusion probabilities  $\pi_{jk|i}$  as in the proposed method, but it induces within-cluster ratio bias when the within-cluster sample sizes are small unlike our method. The expression for  $\hat{S}_{vw}^2$  is more complicated and we refer the reader to Korn and Graubard (2003) for the relevant formula.

The WEE method readily extends to the nested error linear regression model

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + v_i + e_{ij}; \quad e_{ij} \sim_{iid} N(0, \sigma_e^2), \quad v_i \sim_{iid} N(0, \sigma_v^2). \quad (3.13)$$

In this case, the estimating function (3.1) is changed to

$$u_1(y_{ij}, \boldsymbol{\theta}) = \mathbf{x}_{ij} (y_{ij} - \mathbf{x}_{ij}^T \boldsymbol{\beta}), \quad (3.14)$$

(3.2) to

$$u_2(y_{ij}, \boldsymbol{\theta}) = (y_{ij} - \mathbf{x}_{ij}^T \boldsymbol{\beta})^2 - (\sigma_v^2 + \sigma_e^2) \quad (3.15)$$

and (3.3) to

$$u_3(y_{ij}, y_{ik}, \boldsymbol{\theta}) = \left[ z_{ijk} - (\mathbf{x}_{ij} - \mathbf{x}_{ik})^T \boldsymbol{\beta} \right]^2 - 2\sigma_e^2, \quad j \neq k, \quad (3.16)$$

where  $\boldsymbol{\theta}$  is the vector with elements  $\boldsymbol{\beta}$ ,  $\sigma_v^2$  and  $\sigma_e^2$  and  $z_{ijk} = y_{ij} - y_{ik}$ . Explicit solutions to  $\hat{U}_{wt}(\boldsymbol{\theta}) = 0$ ,  $t = 1, 2, 3$  corresponding to (3.14) – (3.16) are obtained as

$$\hat{\boldsymbol{\beta}}_w = \left( \sum_{i \in s} \sum_{j \in s(i)} w_{ij} \mathbf{x}_{ij} \mathbf{x}_{ij}^T \right)^{-1} \left( \sum_{i \in s} \sum_{j \in s(i)} w_{ij} \mathbf{x}_{ij} y_{ij} \right), \quad (3.17)$$

$$\hat{\sigma}_{vw}^2 = \sum_{i \in s} \sum_{j \in s(i)} w_{ij} (y_{ij} - \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}}_w)^2 / \sum_{i \in s} \sum_{j \in s(i)} w_{ij} - \hat{\sigma}_{ew}^2 \quad (3.18)$$

and

$$\hat{\sigma}_{ew}^2 = \sum_{i \in s} w_i \sum_{j < k \in s(i)} w_{jk|i} \left[ z_{ijk} - (\mathbf{x}_{ij} - \mathbf{x}_{ik})^T \hat{\boldsymbol{\beta}}_w \right]^2 / \left( 2 \sum_{i \in s} w_i \sum_{j < k \in s(i)} w_{jk|i} \right). \quad (3.19)$$

## 3.2 Variance estimation

A Taylor linearization sandwich variance estimator of the WEE estimator  $\hat{\boldsymbol{\theta}}_w$  can be obtained along the lines of the variance estimator (2.10), provided the level 2 sampling fraction is small. Let  $\hat{\mathbf{U}}_w(\boldsymbol{\theta})$  be the column vector with components  $\hat{U}_{w1}(\boldsymbol{\theta})$ ,  $\hat{U}_{w2}(\boldsymbol{\theta})$  and  $\hat{U}_{w3}(\boldsymbol{\theta})$  and similarly  $\hat{\mathbf{U}}_{wi}(\boldsymbol{\theta})$  be the column vector with components  $\hat{U}_{w1i}(\boldsymbol{\theta})$ ,  $\hat{U}_{w2i}(\boldsymbol{\theta})$  and  $\hat{U}_{w3i}(\boldsymbol{\theta})$ . Then the linearization variance estimator is given by

$$v_L(\hat{\boldsymbol{\theta}}_w) = (\hat{\mathbf{U}}'_w)^{-1} \left( \sum_{i \in s} w_i^2 \hat{\mathbf{U}}_{wi} \hat{\mathbf{U}}_{wi}^T \right) \left[ (\hat{\mathbf{U}}'_w)^{-1} \right]^T, \quad (3.20)$$

where  $\hat{\mathbf{U}}_{wi}$  and  $\hat{\mathbf{U}}'_w$  denote  $\hat{\mathbf{U}}_{wi}(\boldsymbol{\theta})$  evaluated at  $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_w$  and the first derivative  $\hat{\mathbf{U}}'_w(\boldsymbol{\theta})$  evaluated at  $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_w$ , respectively. Properties of the variance estimator (3.20) are studied through simulation in Section 5.2.

## 4 Weighted log composite likelihood: A unified approach

In this section we propose a unified approach applicable to both linear and generalized linear multi-level models. This approach is based on the concept of composite likelihood which has become popular in the non-survey literature to handle clustered or spatial data (see *e.g.*, Lindsay 1988, Lele and Taper 2002 and Varin, Reid and Firth 2011). A pairwise marginal composite likelihood is obtained by multiplying the likelihood contributions from all the distinct pairs within clusters. Note that the composite likelihood is obtained by pretending the sub-models are independent. When the super-population model holds for the sample, then we can obtain parameter estimators by maximizing the pairwise composite likelihood. Here we extend this approach to handle informative designs by obtaining weighted estimating equations that require only the marginal weights  $w_i$  and  $w_{ji}$  and the pairwise weights  $w_{jk|i}$ , as in Section 3.

The census log pairwise composite likelihood is given by

$$l_C(\boldsymbol{\theta}) = \sum_{i=1}^N \sum_{j < k=1}^{M_i} \log f(y_{ij}, y_{ik} | \boldsymbol{\theta}), \tag{4.1}$$

where  $f(y_{ij}, y_{ik} | \boldsymbol{\theta})$  is the marginal joint density of  $y_{ij}$  and  $y_{ik}$ . We estimate (4.1) by the design-weighted log pairwise composite likelihood

$$l_{wC}(\boldsymbol{\theta}) = \sum_{i \in s} w_i \sum_{j < k \in s(i)} w_{jk|i} \log f(y_{ij}, y_{ik} | \boldsymbol{\theta}) \tag{4.2}$$

which depends only on the first order level 1 and level 2 inclusion probabilities and the second order level 1 probabilities. We then solve the weighted composite score equations

$$\hat{\mathbf{U}}_{wC}(\boldsymbol{\theta}) = \partial l_{wC}(\boldsymbol{\theta}) / \partial \boldsymbol{\theta} = \mathbf{0}, \tag{4.3}$$

obtained from (4.2) to get a weighted composite likelihood estimator,  $\hat{\boldsymbol{\theta}}_{wC}$ , of  $\boldsymbol{\theta}$ . The proposed method is applicable to linear and generalized linear two-level models.

We note that  $\hat{\mathbf{U}}_{wC}(\boldsymbol{\theta})$ , given by (4.3), is a vector of estimating functions with zero expectation with respect to the design and the model, *i.e.*,  $E_m E_p \{ \hat{\mathbf{U}}_{wC}(\boldsymbol{\theta}) \} = \mathbf{0}$ . Using this result, it can be shown that the weighted composite likelihood (WCL) estimator  $\hat{\boldsymbol{\theta}}_{wC}$  of  $\boldsymbol{\theta}$  is design-model consistent as the number of level 2 units in the sample,  $n$ , increases, even when the within cluster sample sizes,  $m_i$ , are small. Details of the proof are given in Yi, Rao and Li (2012). In the non-survey context, we have limited theoretical and empirical evidence that the composite likelihood approach leads to efficient estimators (*e.g.*, Bellio and Varin 2005, Lindsay *et al.* 2011). Our simulation study (Section 5) indicates that the weighted composite likelihood approach performs well in terms of efficiency, even for small within-cluster sample sizes.

In the case of the nested error model (3.13), following Lele and Taper (2002) we can simplify the pairwise composite likelihood approach by replacing the bivariate density function  $f(y_{ij}, y_{ik} | \boldsymbol{\theta})$  by the univariate density functions of  $y_{ij}$  and the difference  $z_{ijk} = y_{ij} - y_{ik}$ . For the mean model (2.2), we have  $y_{ij} \sim N(\mu, \sigma_v^2 + \sigma_e^2)$  and  $z_{ijk} \sim N(0, 2\sigma_e^2)$ . By reparametrizing  $\boldsymbol{\theta} = (\mu, \sigma_v^2, \sigma_e^2)^T$  as  $\boldsymbol{\phi} = (\mu, \sigma^2, \sigma_e^2)^T$

where  $\sigma^2 = \sigma_v^2 + \sigma_e^2$ , we see that the parameters of the two univariate density functions are distinct and the log composite likelihoods corresponding to  $y_{ij}$  and  $z_{ijk}$  are given by

$$l_{wCy}(\boldsymbol{\mu}, \sigma^2) = \sum_{i \in s} w_i \sum_{j \in s(i)} w_{j|i} \log f(y_{ij} | \boldsymbol{\mu}, \sigma^2)$$

and

$$l_{wCz}(\sigma_e^2) = \sum_{i \in s} w_i \sum_{j < k \in s(i)} w_{jkl|i} \log f(z_{ijk} | \sigma_e^2).$$

We then solve the resulting weighted composite score equations

$$\hat{U}_{wCy1}(\boldsymbol{\mu}, \sigma^2) = \partial l_{wCy}(\boldsymbol{\mu}, \sigma^2) / \partial \boldsymbol{\mu} = \sum_{i \in s} w_i \sum_{j \in s(i)} w_{j|i} (y_{ij} - \boldsymbol{\mu}) / \sigma^2 = \mathbf{0},$$

$$\hat{U}_{wCy2}(\boldsymbol{\mu}, \sigma^2) = \partial l_{wCy}(\boldsymbol{\mu}, \sigma^2) / \partial \sigma^2 = \frac{1}{2} \sum_{i \in s} w_i \sum_{j \in s(i)} w_{j|i} \left[ -\frac{1}{\sigma^2} + \frac{(y_{ij} - \boldsymbol{\mu})^2}{\sigma^4} \right] = 0$$

$$\hat{U}_{wCz}(\sigma_e^2) = \partial l_{wCz}(\sigma_e^2) / \partial \sigma_e^2 = \frac{1}{2} \sum_{i \in s} w_i \sum_{j < k \in s(i)} w_{jkl|i} \left[ -\frac{1}{\sigma_e^2} + \frac{z_{ijk}^2}{2\sigma_e^4} \right] = 0$$

to get the weighted composite likelihood (WCL) estimators  $\hat{\boldsymbol{\mu}}_{wC}$ ,  $\hat{\sigma}_{wC}^2$  and  $\hat{\sigma}_{wC}^2$ . The WCL estimators are identical to (3.9) – (3.11) obtained by the weighted estimating equations approach of Section 3.

We now turn to the nested error linear regression model (3.13). We first note that  $y_{ij} \sim N(\mathbf{x}_{ij}^T \boldsymbol{\beta}, \sigma^2)$  where  $\sigma^2 = \sigma_v^2 + \sigma_e^2$ , and  $z_{ijk} = y_{ij} - y_{ik} \sim N\left\{(\mathbf{x}_{ij} - \mathbf{x}_{ik})^T \boldsymbol{\beta}, 2\sigma_e^2\right\}$ . It follows that the weighted composite score equations are given by

$$\begin{aligned} \hat{U}_{wCy1}(\boldsymbol{\beta}, \sigma^2) &= \partial l_{wCy}(\boldsymbol{\beta}, \sigma^2) / \partial \boldsymbol{\beta} \\ &= \sum_{i \in s} w_i \sum_{j \in s(i)} w_{j|i} \mathbf{x}_{ij} (y_{ij} - \mathbf{x}_{ij}^T \boldsymbol{\beta}) = \mathbf{0} \end{aligned}$$

$$\begin{aligned} \hat{U}_{wCy2}(\boldsymbol{\beta}, \sigma^2) &= \partial l_{wCy}(\boldsymbol{\beta}, \sigma^2) / \partial \sigma^2 \\ &= -\frac{1}{2} \sum_{i \in s} w_i \sum_{j \in s(i)} w_{j|i} \left[ \frac{1}{\sigma^2} - \frac{(y_{ij} - \mathbf{x}_{ij}^T \boldsymbol{\beta})^2}{\sigma^4} \right] = 0 \end{aligned}$$

and

$$\begin{aligned} \hat{U}_{wCz}(\sigma_e^2) &= \partial l_{wCz}(\sigma_e^2) / \partial \sigma_e^2 \\ &= -\frac{1}{2} \sum_{i \in s} w_i \sum_{j < k \in s(i)} w_{jkl|i} \left\{ \frac{1}{\sigma_e^2} - \frac{\left[ z_{ijk} - (\mathbf{x}_{ij} - \mathbf{x}_{ik})^T \boldsymbol{\beta} \right]^2}{2\sigma_e^4} \right\} = 0. \end{aligned}$$

The resulting WCL estimators of  $\beta$ ,  $\sigma_v^2$  and  $\sigma_e^2$  are given by

$$\hat{\beta}_{wC} = \left( \sum_{i \in s} \sum_{j \in s(i)} w_{ij} \mathbf{x}_{ij} \mathbf{x}_{ij}^T \right)^{-1} \left( \sum_{i \in s} \sum_{j \in s(i)} w_{ij} \mathbf{x}_{ij} y_{ij} \right),$$

$$\hat{\sigma}_{wC}^2 = \sum_{i \in s} \sum_{j \in s(i)} w_{ij} \left( y_{ij} - \mathbf{x}_{ij}^T \hat{\beta}_{wC} \right)^2 / \sum_{i \in s} \sum_{j \in s(i)} w_{ij},$$

and

$$\hat{\sigma}_{ewC}^2 = \sum_{i \in s} w_i \sum_{j < k \in s(i)} w_{jki} \left[ z_{ijk} - (\mathbf{x}_{ij} - \mathbf{x}_{ik})^T \hat{\beta}_{wC} \right]^2 / \left( 2 \sum_{i \in s} w_i \sum_{j < k \in s(i)} w_{jki} \right).$$

The estimator of  $\sigma_v^2$  is given by  $\hat{\sigma}_{vwC}^2 = \hat{\sigma}_{wC}^2 - \hat{\sigma}_{ewC}^2$ . Again, the WCL estimators  $\hat{\beta}_{wC}$ ,  $\hat{\sigma}_{vwC}^2$  and  $\hat{\sigma}_{ewC}^2$  are identical to (3.17) – (3.19) obtained from the weighted estimating equations approach of Section 3.

The above composite likelihood approach, based on  $y_{ij}$  and  $z_{ijk} = y_{ij} - y_{ik}$ , is not applicable to the linear two-level model given by (2.4) because the parameter vector,  $\theta$ , is not identifiable under the composite likelihood obtained from the  $y_{ij}$  and  $z_{ijk}$ . We need the pairwise method to handle model (2.4).

Marginally,  $(y_{ij}, y_{ik})^T$  is bivariate normal with means  $\mathbf{x}_{ij}^T \beta$  and  $\mathbf{x}_{ik}^T \beta$  and  $2 \times 2$  covariance matrix

$$\Sigma_{i(jk)} = \begin{bmatrix} \sigma_e^2 + \mathbf{x}_{ij}^T \Sigma_v \mathbf{x}_{ij} & \mathbf{x}_{ij}^T \Sigma_v \mathbf{x}_{ik} \\ \mathbf{x}_{ik}^T \Sigma_v \mathbf{x}_{ij} & \sigma_e^2 + \mathbf{x}_{ik}^T \Sigma_v \mathbf{x}_{ik} \end{bmatrix}.$$

It now follows from (4.3) that the weighted composite score equations are given by

$$\beta : \quad \hat{U}_{wC\beta} = \sum_{i \in s} w_i \sum_{j < k \in s(i)} w_{jki} \mathbf{X}_{i(jk)}^T \Sigma_{i(jk)}^{-1} (\mathbf{y}_{i(jk)} - \mathbf{X}_{i(jk)}^T \beta) = \mathbf{0} \tag{4.4}$$

and

$$\tau : \quad \hat{U}_{wC\tau} = \frac{1}{2} \sum_{i \in s} w_i \sum_{j < k \in s(i)} w_{jki} \left[ (\mathbf{y}_{i(jk)} - \mathbf{X}_{i(jk)}^T \beta)^T \Sigma_{i(jk)}^{-1} \frac{\partial \Sigma_{i(jk)}}{\partial \tau_l} \Sigma_{i(jk)}^{-1} (\mathbf{y}_{i(jk)} - \mathbf{X}_{i(jk)}^T \beta) - \text{tr} \left( \Sigma_{i(jk)}^{-1} \frac{\partial \Sigma_{i(jk)}}{\partial \tau_l} \right) \right] = \mathbf{0}, \tag{4.5}$$

$$l = 1, \dots, p(p+1)/2 + 1 = P$$

where  $\mathbf{X}_{i(jk)}$  is the  $2 \times p$  matrix with rows  $\mathbf{x}_{ij}^T$  and  $\mathbf{x}_{ik}^T$ ,  $\mathbf{y}_{i(jk)} = (y_{ij}, y_{ik})^T$  and  $\tau$  is the  $P$ -vector with elements  $\tau_1 = \sigma_e^2$  and the  $p(p+1)/2$  distinct elements of  $\Sigma_v$  denoted by  $\tau_2, \dots, \tau_p$ . We can solve the weighted composite score equations (4.4) and (4.5) iteratively using the Newton-Raphson method or some other iterative method to obtain the WCL estimators  $\hat{\beta}_{wC}$  and  $\hat{\tau}_{wC}$ .

In the special case of the nested error linear regression model (3.13), the census score equations, based on the full census log-likelihood  $l(\theta)$  given by (2.5), can be written in a closed form. The corresponding sample weighted score equations depend only on the level 1 weights  $w_{jli}$  and  $w_{jki}$  and the level 2 weights

$w_i$ , similar to the weighted composite score equations (see the Appendix). The resulting estimators are design-model consistent for  $\theta$ , unlike the estimators based on the weighted pseudo log-likelihood  $l_w(\theta)$  given by (2.7) and (2.8). However, for more complex models, such as two level models with random slopes, the sample weighted score equations will depend on third order and fourth order level 1 inclusion probabilities, unlike the weighted composite score equations (4.3) that depend only on the first order and second order level 1 inclusion probabilities, even for complex multi-level models. We have therefore not included the weighted score equations approach, based on the full census log-likelihood, in the simulation study.

## 5 Simulation study

We conducted a small simulation study on the performance of the proposed WEE estimators under the simple nested error mean model, using  $\mu = 0.5$ ,  $\sigma_v^2 = 0.5$  and  $\sigma_e^2 = 2.0$ . The population consists of  $N = 1,000$  clusters, each containing  $M_i = M = 100$  elements. A two-stage sampling design with  $n = 50$  sample clusters and  $m_i = m = 5$  sample elements from each sample cluster is used. Clusters are selected by simple random sampling, and the elements within clusters by the Rao-Sampford probability proportional to size (PPS) sampling method (Rao 1965 and Sampford 1967) with specified size measures  $z_{ij}$ . The size measures are chosen to reflect different levels of informativeness.

Following Asparouhov (2006), we considered both invariant and non-invariant selections. For invariant selection, the size measure  $z_{ij}$  depends only on the level 1 errors and is invariant across clusters.

In particular, we let

$$z_{ij} = \left( 1 + \exp \left\{ -0.5 \left[ e_{ij} / \alpha + e_{ij}^* (1 - \alpha^{-2})^{1/2} \right] \right\} \right)^{-1}, \quad (5.1)$$

where  $e_{ij}^*$  is independent of  $e_{ij}$  but with the same distribution,  $N(0, \sigma_e^2 = 2.0)$ . For non-invariant selection, the size measure  $z_{ij}$  depends on both level 1 and level 2 errors and hence non-invariant across clusters. In particular, we replace  $e_{ij}$  and  $e_{ij}^*$  in (3.7) by  $v_i + e_{ij}$  and  $v_i^* + e_{ij}^*$  respectively, where  $v_i^*$  is independent of  $v_i$  but with the same distribution  $N(0, \sigma_v^2 = 0.5)$ . We considered four values of  $\alpha$  in (5.1):  $\alpha = 1, 2, 3, \infty$ , where  $\alpha = \infty$  corresponds to non-informative sampling within each cluster,  $\alpha = 1$  corresponds to the most informative sampling and informativeness decreases as  $\alpha$  increases.

We used the design-model (*pm*) approach to simulate  $R = 1,000$  samples for each specified  $\alpha$  and separately for invariant and non-invariant selections. Under this approach, we generated a population with  $N = 1,000$  and  $M_i = M = 100$  from the model and then selected a two-stage sample of elements as specified above. The two-step process was repeated  $R = 1,000$  times to simulate 1,000 samples.

### 5.1 Performance of estimators

From each sample, we computed the estimates of  $\mu, \sigma_v^2$  and  $\sigma_e^2$  using REML, weighted scaling methods A and A1, the proposed WEE method and the alternative method of Korn and Graubard

(abbreviated KG). Biases and variances of the estimators were computed from the 1,000 estimates. Performance of alternative estimators is judged using two performance measures: Bias ratio = BR = (Bias)/(square root of variance) and relative root mean squared error = RRMSE = (square root of MSE)/(true parameter value). Tables 5.1, 5.2 and 5.3 respectively report the BR values of the estimators of  $\mu$ ,  $\sigma_v^2$  and  $\sigma_e^2$ . RRMSE values of the estimators of  $\mu$ ,  $\sigma_v^2$  and  $\sigma_e^2$  are reported in Tables 5.4, 5.5 and 5.6 respectively.

**Table 5.1**  
**Bias ratio (%) of estimators of  $\mu$**

$\alpha$	Invariant			Non-invariant		
	REML	A	A1/WEE/KG	REML	A	A1/WEE/KG
1	346.5	80.2	2.2	370.9	83.9	3.0
2	167.7	40.1	0.3	172.3	45.3	6.1
3	114.3	30.7	4.5	114.9	30.8	4.8
$\infty$	2.0	2.5	2.1	-1.5	-2.4	-2.2

Table 5.1 reports bias ratio (%) of the estimators of  $\mu$  based on REML, weight-scaling methods A and A1, KG and WEE. Note that in the case of  $\mu$ , estimators A1, KG and WEE (WCL) are identical. Results in Table 5.1 show that BR is similar for invariant and non-invariant selections and that BR of REML and A decrease as  $\alpha$  increases. Further, REML leads to large bias under informative sampling, even for  $\alpha = 3$ ; for example, BR for REML ranges from 114% to 346% under invariant selection. Method A also leads to significant BR under informative sampling; for example BR for A ranges from 30.8% to 83.9% under non-invariant selection. On the other hand, BR of WEE, A1 and KG does not depend on  $\alpha$  and it is small ( $|BR| < 6\%$ ). Under non-informative sampling, REML performs well as expected ( $|BR| < 3\%$ ).

Turning to the estimation of  $\sigma_v^2$ , we first note that the proportion of times the estimate of  $\sigma_v^2$  is negative is zero in the simulations for all four values of  $\alpha$  and for all the estimation methods (REML, A, A1, WEE and KG). Table 5.2 reports BR values of the estimators of  $\sigma_v^2$ . It shows that the BR of REML is not affected by  $\alpha$  under invariant selection, but is affected under non-invariant selection. In the latter case, REML leads to serious underestimation for  $\alpha = 1$  (BR = -49%) but  $|BR|$  decreases as  $\alpha$  increases. Table 5.2 also shows that methods A and A1 do not perform well under informative sampling (BR ranging from 16% to 60%). KG did not perform well for  $\alpha = 1$  (BR=33% under invariant selection and BR = 24% under non-invariant selection). On the other hand, WEE performs well for all values of  $\alpha$  (BR ranging from -4% to -13%) although underestimation is consistent across values of  $\alpha$ .

Table 5.3 reports BR values of the estimators of  $\sigma_e^2$ . It shows that BR values are similar for invariant and non-invariant selections, as in the case of  $\mu$ . REML and KG lead to serious underestimation when  $\alpha = 1$  (BR = -107% for REML and BR = -71% for KG under invariant selection), but  $|BR|$  decreases as  $\alpha$  increases and becomes negligible for  $\alpha = \infty$ . Estimators A and A1 perform poorly for all values of  $\alpha$ .

including  $\alpha = \infty$ . On the other hand, WEE performs well for all values of  $\alpha$  with  $|\text{BR}| < 8\%$ . It appears that the instability introduced by the scale factor (2.9) might have contributed to the large  $|\text{BR}|$  for methods A and A1 even for the case of non-informative sampling ( $\alpha = \infty$ ).

**Table 5.2****Bias ratio (%) of estimators of  $\sigma_v^2$** 

$\alpha$	REML	A	A1	WEE	KG
<b>Invariant Selection</b>					
1	0.6	59.5	59.3	-8.5	33.2
2	0.5	24.5	26.3	-10.0	8.0
3	-3.4	16.1	18.2	-13.6	0.4
$\infty$	-0.1	14.8	17.1	-8.9	0.6
<b>Non-invariant Selection</b>					
1	-49.0	50.1	58.9	-4.4	24.0
2	-10.9	24.6	28.7	-7.0	7.1
3	-4.0	20.0	22.7	-7.8	4.6
$\infty$	-1.3	12.8	13.9	-13.3	-1.6

**Table 5.3****Bias ratio (%) of estimators of  $\sigma_e^2$** 

$\alpha$	REML	A	A1	WEE	KG
<b>Invariant Selection</b>					
1	-106.9	-118.4	-66.9	2.4	-71.2
2	-22.7	-43.6	-34.3	2.1	-16.5
3	-9.4	-31.7	-28.4	2.9	-6.5
$\infty$	-0.4	-21.8	-23.8	0.3	0.4
<b>Non-invariant Selection</b>					
1	-115.3	-131.3	-79.6	-6.9	-82.6
2	-30.4	-51.1	-43.3	-7.6	-23.9
3	-12.5	-34.9	-32.2	-2.3	-10.3
$\infty$	1.1	-20.2	-21.8	2.6	1.6

**Table 5.4**  
**Relative root mean squared error (%) of estimators of  $\mu$**

$\alpha$	Invariant			Non-invariant		
	REML	A	A1/WEE/KG	REML	A	A1/WEE/KG
1	93.3	35.9	29.4	92.5	35.4	29.2
2	51.6	29.3	27.8	52.8	30.4	28.9
3	40.5	28.2	27.5	40.8	28.7	28.1
$\infty$	25.8	26.1	26.5	26.6	27.3	27.7

*Relative root mean squared error*

Table 5.4 shows that the RRMSE (%) values for estimators of  $\mu$  are similar for invariant and non-invariant selections and that RRMSE of REML and A decrease as  $\alpha$  increases. For informative sampling with  $\alpha = 1$ , RRMSE for REML is large relative to RRMSE for WEE (A1 and KG) due to large BR. For example, RRMSE=93% for REML compared to RRMSE=29% for WEE. As expected, REML has the smallest RRMSE under non-informative sampling, but the increase in RRMSE for the other methods is quite small. Also, RRMSE of WEE (A1 and KG) depends on  $\alpha$ .

**Table 5.5**  
**Relative root mean squared error (%) of estimators of  $\sigma_v^2$**

$\alpha$	REML	A	A1	WEE	KG
	Invariant Selection				
1	36.5	47.3	51.1	43.6	43.8
2	37.1	39.7	41.1	40.5	39.5
3	36.3	37.3	38.7	39.5	37.8
$\infty$	35.8	36.9	38.1	38.7	37.2
Non-invariant Selection					
1	36.7	44.6	52.6	43.4	41.5
2	35.6	37.9	40.4	39.3	37.7
3	37.0	38.7	40.4	40.2	38.8
$\infty$	36.6	37.2	38.0	39.0	37.8

Turning to RRMSE of estimators of  $\sigma_v^2$ , Table 5.5 shows that REML performs well for all  $\alpha$  under invariant selection due to small BR in this case. We also note that KG and WEE are comparable in terms of RRMSE for all values of  $\alpha$ . Table 5.5 also shows that A and A1 lead to somewhat larger RRMSE for  $\alpha = 1$ : 51% for A1 and 47% for A under invariant selection compared to 44% for WEE.

**Table 5.6****Relative root mean squared error (%) of estimators of  $\sigma_e^2$** 

$\alpha$	REML	A	A1	WEE	KG
<b>Invariant Selection</b>					
1	13.5	14.5	12.8	13.9	12.9
2	9.7	10.4	10.4	11.0	10.0
3	9.5	10.0	10.1	10.7	9.8
$\infty$	10.1	10.3	10.5	11.1	10.3
<b>Non-invariant Selection</b>					
1	13.7	14.8	12.9	13.2	13.0
2	10.0	10.9	10.9	11.3	10.3
3	9.7	10.4	10.7	11.2	10.2
$\infty$	10.3	10.6	10.8	11.4	10.7

Table 5.6 gives RRMSE values of the estimators of  $\sigma_e^2$  and we note that the values are similar for invariant and non-invariant selections. It also shows that RRMSE values are comparable for methods WEE, A, A1 and KG even though in terms of bias ratio A, A1 and KG performed poorly relative to WEE. This is due to larger variance for WEE compared to other methods. For example, in the case of invariant selection and  $\alpha = 1$  we have the following variances for WEE, KG and REML: 0.0771, 0.0438 and 0.0339 with corresponding bias ratios (%) from Table 5.3: 2.4, -71.2, and -106.9.

## 5.2 Performance of variance estimator

We now report some simulation results on the relative bias of the linearization variance estimator (3.12) of the WEE (WCL) estimator  $\hat{\theta}_w$ . We first repeated the two-step process  $R_1 = 2,000$  times and computed  $v_L^{(r)}(\hat{\theta}_w)$  from each two-stage sample  $r = 1, \dots, 2,000$ . The averages of the diagonal elements of  $E\{v_L(\hat{\theta}_w)\} \approx v_L(\hat{\theta}_w) = R_1^{-1} \sum_{r=1}^{R_1} v_L^{(r)}(\hat{\theta}_w)$  are denoted by  $\bar{v}_L(\hat{\mu}_w)$ ,  $\bar{v}_L(\hat{\sigma}_{vw}^2)$  and  $\bar{v}_L(\hat{\sigma}_{ew}^2)$  respectively. We then generated  $R_2 = 10,000$  independent samples and computed the empirical mean squared error

(MSE) of the three estimators  $\hat{\mu}_w$ ,  $\hat{\sigma}_{vw}^2$  and  $\hat{\sigma}_{ew}^2$ . We have  $MSE(\hat{\mu}_w) \approx R_2^{-1} \sum_{r=1}^{R_2} (\hat{\mu}_w^{(r)} - \mu)^2$  where  $\hat{\mu}_w^{(r)}$  is the estimate of  $\mu$  from the  $r$ -th simulated sample, and similar expressions for  $MSE(\hat{\sigma}_{vw}^2)$  and  $MSE(\hat{\sigma}_{ew}^2)$ .

The relative bias of  $v_L(\hat{\mu}_w)$  is calculated as

$$RB\{v_L(\hat{\mu}_w)\} = [\bar{v}_L(\hat{\mu}_w)/MSE(\hat{\mu}_w)] - 1$$

and similarly  $RB\{v_L(\hat{\sigma}_{vw}^2)\}$  and  $RB\{v_L(\hat{\sigma}_{ew}^2)\}$  were calculated. Table 5.7 reports the RB values of the three variance estimators for invariant and non-invariant selections and  $\alpha = 1, 2, 3, \infty$ . It is clear from Table 5.7 that the linearization variance estimator performs well over all combinations with  $|RB| < 10\%$ .

**Table 5.7**  
**Relative bias (%) of variance estimators**

$\alpha$	$v_L(\hat{\mu}_w)$	$v_L(\hat{\sigma}_{vw}^2)$	$v_L(\hat{\sigma}_{ew}^2)$
<b>Invariant Selection</b>			
1	-3.0	-6.2	-7.5
2	-5.2	-4.5	-3.1
3	-1.3	-3.8	-1.8
$\infty$	-0.9	-2.5	-2.0
<b>Non-invariant Selection</b>			
1	-3.8	-8.3	-4.2
2	-4.5	-5.8	-7.3
3	-4.3	-4.6	-5.7
$\infty$	-2.4	-2.7	-2.9

## 6 Concluding remarks

In this paper, we have proposed a unified weighted composite likelihood (WCL) approach for two-level models to make inferences from complex survey data. The proposed WCL methods are asymptotically valid even when the sample sizes within sampled clusters (level 1 units) are small, unlike some of the existing methods, but knowledge of the joint inclusion probabilities within sampled clusters is required. Often it may be possible to treat the sample within clusters as drawn with replacement because of small sampling fractions within clusters. Also, excellent approximations to joint inclusion probabilities, depending only on the marginal inclusion probabilities, are also available when the sampling fractions are not small (Haziza *et al.* 2008). We plan to study the accuracy of such approximations in a future study.

Simulation studies on the performance of the WCL estimators (4.5) and (4.6) for two-level models (2.3), based on the pairwise method, will also be conducted.

Composite likelihood methods are mostly used when the full likelihood is complex. Our development in the survey sampling context demonstrates that the full likelihood method with weights is not feasible for multi-level models whereas the weighted composite likelihood method facilitates valid inferences even when the cluster sample sizes are small.

## 7 Acknowledgements

We thank two referees and the associate editor for constructive comments and suggestions.

## Appendix

### Weighted score equations: nested error linear regression model

For the nested error linear regression model (2.3), an explicit form for the census full log-likelihood is obtained using the explicit form for the covariance matrix  $\mathbf{V}_i$  of  $\mathbf{y}_i = (y_{i1}, \dots, y_{iM_i})^T$ . We have  $\mathbf{V}_i^{-1} = \sigma_e^{-2} [\mathbf{I}_i - \sigma_v^2 / \lambda_i \mathbf{1}_i \mathbf{1}_i^T]$ , where  $\lambda_i = \sigma_e^2 + M_i \sigma_v^2$ ,  $\mathbf{I}_i$  is the  $M_i \times M_i$  identity matrix and  $\mathbf{1}_i$  is the  $M_i \times 1$  unit vector. Using the expression for  $\mathbf{V}_i^{-1}$ , the census score equations are obtained as

$$\boldsymbol{\beta}: \left[ \sum_{i=1}^N \sum_{j=1}^{M_i} \mathbf{x}_{ij} y_{ij} - \sigma_v^2 \sum_{i=1}^N \lambda_i^{-1} \left( \sum_{j=1}^{M_i} \sum_{k=1}^{M_i} \mathbf{x}_{ij} y_{ik} \right) \right] - \left[ \sum_{i=1}^N \sum_{j=1}^{M_i} \mathbf{x}_{ij} \mathbf{x}_{ij}^T - \sigma_v^2 \sum_{i=1}^N \lambda_i^{-1} \left( \sum_{j=1}^{M_i} \sum_{k=1}^{M_i} \mathbf{x}_{ij} \mathbf{x}_{ik}^T \right) \right] \boldsymbol{\beta} = 0 \quad (\text{A.1})$$

$$\sigma_v^2: \sum_{i=1}^N \lambda_i^{-2} \left[ \sum_{j=1}^{M_i} \sum_{k=1}^{M_i} (y_{ij} - \mathbf{x}_{ij}^T \boldsymbol{\beta})(y_{ik} - \mathbf{x}_{ik}^T \boldsymbol{\beta}) \right] - \sum_{i=1}^N \lambda_i^{-1} M_i = 0 \quad (\text{A.2})$$

$$\begin{aligned} \sigma_e^2: & \sum_{i=1}^N \sum_{j=1}^{M_i} (y_{ij} - \mathbf{x}_{ij}^T \boldsymbol{\beta})^2 + \sum_{i=1}^N (M_i \sigma_v^4 \lambda_i^{-2} - 2 \sigma_v^2 \lambda_i^{-1}) \sum_{j,k=1}^{M_i} (y_{ij} - \mathbf{x}_{ij}^T \boldsymbol{\beta})(y_{ik} - \mathbf{x}_{ik}^T \boldsymbol{\beta}) \\ & - \sigma_e^2 \sum_{i=1}^N (1 - \sigma_v^2 \lambda_i^{-1}) M_i = 0 \end{aligned} \quad (\text{A.3})$$

From (A.1), we obtain weighted score equations

$$\begin{aligned} \boldsymbol{\beta}: & \sum_{i \in \mathcal{S}} w_i \sum_{j \in \mathcal{S}(i)} w_{j|i} \mathbf{x}_{ij} y_{ij} - \sigma_v^2 \sum_{i \in \mathcal{S}} w_i \lambda_i^{-1} \left( \sum_{j \in \mathcal{S}(i)} \sum_{k \in \mathcal{S}(i)} w_{j|k|i} \mathbf{x}_{ij} y_{ik} \right) \\ & - \left[ \sum_{i \in \mathcal{S}} w_i \sum_{j \in \mathcal{S}(i)} w_{j|i} \mathbf{x}_{ij} \mathbf{x}_{ij}^T - \sigma_v^2 \sum_{i \in \mathcal{S}} w_i \lambda_i^{-1} \left( \sum_{j \in \mathcal{S}(i)} \sum_{k \in \mathcal{S}(i)} w_{j|k|i} \mathbf{x}_{ij} \mathbf{x}_{ik}^T \right) \right] \boldsymbol{\beta} = 0 \end{aligned} \quad (\text{A.4})$$

where  $w_{j|i} = w_{j|i}$ . Note that the cluster sizes  $M_i$  for  $i \in s$  are assumed to be known. One should not replace  $M_i$  by its estimate  $\sum_{j \in s(i)} w_{j|i}$  because it includes ratio bias due to small within cluster sample sizes. The estimating equation (A.4) is design-unbiased for the census equation (A.1).

Turning to the weighted score equation for  $\sigma_v^2$ , we obtain from (A.2)

$$\sigma_v^2 : \sum_{i \in s} w_i \lambda_i^{-2} \left[ \sum_{j \in s(i)} \sum_{k \in s(i)} w_{jk|i} (y_{ij} - \mathbf{x}_{ij}^T \boldsymbol{\beta})(y_{ik} - \mathbf{x}_{ik}^T \boldsymbol{\beta}) \right] - \sum_{i \in s} w_i \lambda_i^{-1} \sum_{j \in s(i)} w_{j|i} = 0 \quad (\text{A.5})$$

The estimating equation (A.5) is unbiased for (A.2). Finally, the weighted score equation for  $\sigma_e^2$  is obtained from (A.3) as

$$\begin{aligned} \sigma_e^2 : & \sum_{i \in s} w_i \sum_{j \in s(i)} w_{j|i} (y_{ij} - \mathbf{x}_{ij}^T \boldsymbol{\beta})^2 + \sum_{i \in s} w_i (M_i \sigma_v^4 \lambda_i^{-2} - 2\sigma_v^2 \lambda_i^{-1}) \sum_{j,k \in s(i)} w_{jk|i} (y_{ij} - \mathbf{x}_{ij}^T \boldsymbol{\beta})(y_{ik} - \mathbf{x}_{ik}^T \boldsymbol{\beta}) \\ & - \sigma_e^2 \sum_{i \in s} w_i (1 - \sigma_v^2 \lambda_i^{-1}) \sum_{j \in s(i)} w_{j|i} = 0 \end{aligned} \quad (\text{A.6})$$

It follows from (A.4) – (A.6) that the weighted score equations depend only on the first order weights  $w_i$  and  $w_{j|i}$  and the second order weights  $w_{jk|i}$  in the special case of a nested error linear regression model.

## References

- Asparouhov, T. (2006). Generalized multi-level modeling with sampling weights. *Communications in Statistics - Theory and Methods*, 35, 439-460.
- Beaumont, J.-F., and Charest, A.-S. (2010). Bootstrap variance estimation with survey data when estimating model parameters. Unpublished report (courtesy of the authors).
- Bellio, R., and Varin, C. (2005). A pairwise likelihood approach to generalized linear models with crossed random effects. *Statistical Modelling*, 3, 217-227.
- Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika*, 73, 43-56.
- Grilli, L., and Pratesi, M. (2004). Weighted estimation in multilevel 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.
- Korn, E.L., and Graubard, B.I. (2003). Estimating variance components using survey data. *Journal of the Royal Statistical Society B*, 65, 175-190.
- Kovacevic, M.S., Rong, H. and You, Y. (2006). Bootstrapping for variance estimation in multi-level models fitted to survey data. *Proceedings of ASA Section on Survey Research Methods*, American Statistical Association, 3260-3269.

- Lele, S., and Taper, M.L. (2002). A composite likelihood approach to (co)variance components estimation. *Journal of Statistical Planning and Inference*, 103, 117-125.
- Lindsay, B.G. (1988). Composite likelihood methods. In *Statistical Inference from Stochastic Processes*, (Ed. N.U. Prabhu), Providence: American Mathematical Society, 221-239.
- Lindsay, B.G., Yi, G.Y. and Sun, J. (2011). Issues and strategies in the selection of composite likelihoods. *Statistica Sinica*, 21, 71-105.
- Muthén, L.K., and Muthén, B.O. (1998-2005). *Mplus User's Guide*. 3<sup>rd</sup> ed. Los Angeles, CA: Muthén & Muthén.
- Pfeffermann, D., and Sverchkov, M. (2003). Fitting generalized linear models under informative sampling. In *Analysis of Survey Data*, (Eds. R. Chambers and C.J. Skinner) 175-196, Wiley, Chichester.
- 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 B*, 60, 23-56.
- Rabe-Hesketh, S., and Skrondal, A. (2006). Multilevel modeling of complex survey data. *Journal of the Royal Statistical Society A*, 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., and Roberts, G. (1998). Discussion on the papers by Firth and Bennett and Pfeffermann *et al.* *Journal of the Royal Statistical Society B*, 60, 50-51.
- Rao, J.N.K., Wu, C.F.J. and Yue, K. (1992). Some recent work on resampling methods for complex surveys. *Survey Methodology*, 18, 209-217.
- Rao, J.N.K., Hidirolou, M., Yung, W. and Kovacevic, M. (2010). Role of weights in descriptive and analytical inferences from survey data: An overview. *Journal of the Indian Society of Agricultural Statistics*, 64, 129-135.
- Sampford, M.R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika*, 54, 499-513.
- Varin, C., Reid, N. and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, 21, 5-42.
- Yi, G.Y., Rao, J.N.K. and Li, H. (2012). A weighted composite likelihood approach for analysis of survey data under two level models. Available on request to [jrao@math.carleton.ca](mailto:jrao@math.carleton.ca).