# Modelling multilevel data under complex sampling designs: an empirical likelihood approach[⋆]

Melike Oğuz-Alper[a], Yves G. Berger[b,*]

[a] Statistics Norway, Postboks 2633 St. Hanshaugen, NO-0131 Oslo, Norway
[b] University of Southampton, SO17 1BJ, United Kingdom

ARTICLE INFO

*Keywords*:
design-based inference
generalised estimating equation
two-stage sampling
uniform correlation structure
regression coefficient
unequal inclusion probability

ABSTRACT

Data used in social, behavioural, health or biological sciences may have a hierarchical structure due to the population of interest or the sampling design. Multilevel or marginal models are often used to analyse such hierarchical data. These data are often selected with unequal probabilities from a clustered and stratified population. An empirical likelihood approach for the regression parameters of a multilevel model is proposed. It has the advantage of taking into account of the sampling design. This approach can be used for point estimation, hypothesis testing and confidence intervals for the sub-vector of parameters. It provides asymptotically valid inference for small and large sampling fraction. The simulation study shows the advantages of the empirical likelihood approach over alternative parametric approaches. The approach proposed is illustrated using the Programme for International Student Assessment (PISA) survey data.

## 1. Introduction

Multilevel (Goldstein, 1986) or marginal models (Diggle, Heagerty, Liang and Zeger, 2002) are often used to analyse hierarchical data. Sample data are often selected from multi-stage sampling designs involving unequal probabilities at the first stage of the selection. Ignoring the selection probabilities may result in invalid inference, when these probabilities are associated with the model outcome variable after conditioning on the model covariates (Pfeffermann, Skinner, Holmes, Goldstein and Rasbash, 1998).

With single level regression models, sampling weights can be taken into account by using the *pseudo-likelihood* approach (Binder, 1983; Binder and Patak, 1994; Skinner, 1989), where the population is fixed and the observations are assumed independent. It is not straightforward to implement this approach with multilevel models, because the observations within higher levels of the hierarchy are not marginally independent. In this case, population totals cannot be written as a single summation of the individual units (Grilli and Pratesi, 2004).

Pfeffermann et al. (1998) proposed a *weighted iterative generalised least squares* (WIGLS) algorithm to estimate multilevel regression parameters under two-stage sampling design. This procedure can be computationally intensive (Kovačević and Rai, 2003). A weighted *multilevel pseudo-likelihood* approach was proposed by Asparouhov (2006) and Rabe-Hesketh and Skrondal (2006). Their point estimator is approximately unbiased if the within cluster sample sizes are large enough, which is may not be the case in practice. Skinner and Vieira (2007) proposed a *weighted generalised estimating equation* (WGEE) approach, by incorporating the sampling weights into generalised estimating equation (Liang and Zeger, 1986). Skinner and Vieira (2007, p.5) noticed that the WIGLS estimator and the WGEE estimator are almost identical under a working uniform correlation structure.

Rao, Verret and Hidiroglou (2013) proposed a *weighted composite likelihood*, which can be used for point estimation of two-level models. Design-based estimators are obtained by solving sample composite score equations, based upon the weighted pairwise composite likelihood function. The estimators are both design and model consistent. This approach is asymptotically valid even when the cluster sample sizes are small, unlike weighted multilevel

[⋆] **This is the pre-peer reviewed version of the following article: Oğuz-Alper and Berger (2020)**
https://doi.org/10.1016/j.csda.2019.106906
There is a supplementary material for this article that can be found online at https://doi.org/10.1016/j.csda.2019.106906.
Computational codes concerning the empirical likelihood approach proposed in this article can be found online at http://yvesberger.co.uk.

[*] Corresponding author
✉ Y.G.Berger@soton.ac.uk ( Yves G. Berger)
http://yvesberger.co.uk ( Yves G. Berger)
ORCID(s): 0000-0002-9128-5384 ( Yves G. Berger)

pseudo-likelihood (Asparouhov, 2006; Rabe-Hesketh and Skrondal, 2006). However, joint inclusion probabilities have to be known for point estimation with weighted composite likelihood. The approach proposed does not rely on joint inclusion probabilities

Pfeffermann, Moura and Silva (2006) and Kim, Park and Lee (2017) proposed fully parametric approaches. These approaches are based on parametric assumption which are not needed under the approach proposed. Kim et al. (2017) considered an expectation-maximisation (EM) algorithm based on the assumption that distribution of the estimator of the random effect is approximately normal. Pfeffermann et al.'s (2006) approach is based on a model holding for the sample data in terms of the population model and the selection probabilities for each stage of the sample selection. The sample model is fitted by using MCMC. The validity of this method depends on the sample model being correctly specified. It is well known that parametric approaches may perform badly when the model assumptions they are based upon are not met. The primary focus is on non-parametric (design-based) approaches. A semi-parametric (model-design-based) approach is also considered in §6.

Under a design-based approach (Neyman, 1938), the sampling distribution is specified by the sampling design. Population-level information can be accommodated within the approach proposed. The design-consistent point estimator proposed is the solution to generalised estimating equation (see §3). Consider the empirical likelihood confidence intervals based on a pivotal empirical log-likelihood ratio function. These intervals do not rely on re-sampling, linearisation, variance estimation or design effect. These are key advantages over alternative approaches. For example, the pseudoempirical log-likelihood ratio function is not pivotal and needs to be adjusted by design effects based on variance estimates. Tan and Wu's (2015) empirical log-likelihood ratio function is only pivotal when the parameter of interest is unidimensional.

Consider a multidimensional parameter of multilevel regression models. Profiling consists in maximising the empirical likelihood function over the parameters which are not of interest. The resulting profile empirical log-likelihood ratio function is pivotal, which allows constructing confidence intervals for each components of the parameter and testing the significance of sub-parameters. The pseudoempirical likelihood approaches and Tan and Wu's (2015) approach are limited to the unidimensional case, and there is no extension for the multidimensional case. There is no general multidimensional theory on profiling for the pseudoempirical likelihood likelihood approach. Profiling is the the key property of the approach proposed for multilevel and marginal regression model parameters. A comparison of empirical likelihood and pseudoempirical likelihood can be found in Berger (2018b,c).

Like bootstrap, the approach proposed does not require variance estimation, design effects and linearisation. Bootstrap can be very computationally intensive for hierarchical data, because many replicates may be needed for bootstrap confidence intervals. Furthermore, bootstrap relies on small sampling fractions, which is not required with the approach proposed (see §6). It is also less computer intensive than the bootstrap. The theoretical properties of bootstrap under complex sampling are often conjectured, only supported by simulation studies or limited to simple random sampling. The properties of the approach proposed has the advantage of being based on an asymptotic framework. It also allows for large sampling fractions.

It is assumed that the model and the design have the same hierarchical structure; that is, the model and design clusters do not overlap. This is a standard assumption made by Pfeffermann et al. (1998); Skinner and Vieira (2007) and Rao et al. (2013). The empirical likelihood function is defined at *primary sampling units* (PSUs) level. The PSU-level sampling fraction can negligible or large.

Standard confidence intervals of estimators based on estimating equations, often rely on *linearised sandwich variance estimator* (e.g. Binder, 1983; Kovačević and Rai, 2003; Pfeffermann et al., 1998; Rao et al., 2013; Skinner and Vieira, 2007) or bootstrap (Grilli and Pratesi, 2004). Standard confidence intervals may have poor coverages, when the variance estimators are biased or unstable. This may be the case when the sample size is not large enough or with outlying values. Even under normality, heteroskedasticity may affect the coverage of standard confidence intervals (Owen, 1991).

In §2, two-stage sampling designs and population-level information are introduced. In §3, the multilevel model considered and the parameter of interest are defined. In §4, the WGEE point estimator is defined. In §5, the empirical likelihood approach proposed is described. An extension for large sampling fractions can be found in §6. In §7, the performance of the empirical likelihood confidence interval is compared with alternative parametric approaches. In §8, the approach proposed is applied to the Programme for International Student Assessment (PISA) survey data (OECD, 2006).

## 2. Two-stage sampling design and population-level information

Let $U$ be a finite population comprised of $N$ disjoint finite *primary sampling units* (PSUs) $U_i$ of sizes $K_i$, with $i = 1, \dots, N$. Suppose that $U$ is stratified into a finite number $H$ of strata denoted by $U_1, \dots, U_H$, such that $\cup_{h=1}^{H} U_h = U$ and $\Sigma_{h=1}^{H} N_h = N$, where $N_h$ denotes the number of PSUs within $U_h$. Let $S_h$ be the sample of $U_i$, selected *without replacement* with unequal probabilities $\pi_i$ from $U_h$. Let $n_h$ denote the sample size of $U_h$. The overall sample of PSUs is $S = \cup_{h=1}^{H} S_h$. Let $S_i$ be the sample of *secondary sampling units* (SSUs), of size $k_i$ selected with conditional probabilities $\pi_{j|i}$ within the $i$th PSU selected at the first stage, with $j = 1, \dots, k_i$. Let $K_i$ denote the size of $U_i$. Let $\boldsymbol{v}_{ij}$ be the vector of variables associated with unit $j \in U_i$.

Consider some known population parameter $\boldsymbol{\varphi}_N$ which is the solution to the estimating equation (Chaudhuri, Handcock and Rendall, 2008)

$$\sum_{i \in U} \sum_{j \in U_i} \mathbf{f}(\boldsymbol{\varphi}, \boldsymbol{v}_{ij}) = \mathbf{0} \cdot \tag{1}$$

For example, $\boldsymbol{\varphi}_N$ could be a vector of means, ratios or quantiles. The vector $\boldsymbol{\varphi}_N$ will be treated as a vector of constants, not as a parameter to estimate. For simplicity, $\mathbf{f}(\boldsymbol{\varphi}_N, \boldsymbol{v}_{ij})$ is replaced by $\mathbf{f}_{ij}$ in what follows. The $\mathbf{f}_{ij}$ are often called auxiliary information in the sampling literature.

The asymptotic framework considered is based on an infinite nested sequence of sampling designs, a sequence of finite populations and an associated sequence of samples (Isaki and Fuller, 1982). Assume that $n \to \infty$, where $n$ is the number of PSUs sampled. The sizes $K_i$ are assumed asymptotically bounded. Thus, $k_i$ is finite, unlike the weighted multilevel pseudo-likelihood approach (Asparouhov, 2006; Rabe-Hesketh and Skrondal, 2006). The number of strata $H$, $n_h/n$ and $N_h/N$ are fixed constants that do not vary as $n \to \infty$. The sampling fraction $n/N$ can be negligible or large. The extension to large sampling fractions can be found in §6.

## 3. Multilevel model

Let $y_{ij}$ be the values of a variable of interest and $\boldsymbol{x}_{ij}$ be the vector of values of $b$ explanatory variables. The variables $y_{ij}$ and $\boldsymbol{x}_{ij}$ are associated with the $j$th unit within the $i$th cluster, where $j = 1, \dots, K_i$ and $i = 1, \dots, N$. The variables $y_{ij}$ and $\boldsymbol{x}_{ij}$ are considered to be part of a vector $\boldsymbol{v}_{ij}$; that is, $\boldsymbol{v}_{ij} = (y_{ij}, \boldsymbol{x}_{ij}^{\top}, \dots)^{\top}$. Consider the multilevel model

$$y_{ij} = \boldsymbol{x}_{ij}^{\top} \boldsymbol{B} + \epsilon_{ij}, \quad \text{where } \epsilon_{ij} := u_i + e_{ij} \cdot \tag{2}$$

Here, $u_i$ and the $e_{ij}$ are independent random variables with means zero and variances $\sigma_u^2$ and $\sigma_e^2$ respectively. The response variables $y_{ij}$ are conditionally independent given the random effects $u_i$ and marginally correlated within cluster $i$. This implies that the variance of $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \dots, \epsilon_{iK_i})^{\top}$, with respect to (2), is

$$\boldsymbol{\Sigma}_i = \sigma_e^2 \, \boldsymbol{I}_{K_i} + \sigma_u^2 \, \mathbf{1}_{K_i} \mathbf{1}_{K_i}^{\top},$$

where $\boldsymbol{I}_{K_i}$ is the $K_i \times K_i$ identity matrix and $\mathbf{1}_{K_i}$ is the $K_i \times 1$ vector of ones. The approach proposed can be extended to more complex multilevel models with random slopes and/or complex correlation structures. For simplicity and without loss of generality, the focus will be upon the model (2).

Let the finite population parameter $\boldsymbol{\beta}_N$ be the *generalised least square* predictor of $\boldsymbol{B}$. The parameter $\boldsymbol{\beta}_N$ is the solution to the population *generalised estimating equation* (Liang and Zeger, 1986)

$$\boldsymbol{G}(\boldsymbol{\beta}) := \sum_{i \in U} \boldsymbol{g}_i(\boldsymbol{\beta}) = \mathbf{0}_b, \tag{3}$$

where

$$\begin{aligned}
\boldsymbol{g}_i(\boldsymbol{\beta}) &:= \boldsymbol{X}_i^{\top} \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{\beta}), \\
\boldsymbol{y}_i &:= (y_{i1}, \dots, y_{iK_i})^{\top}, \\
\boldsymbol{X}_i &:= (\boldsymbol{x}_{i\cdot}^{(1)}, \cdots, \boldsymbol{x}_{i\cdot}^{(b)}), \\
\boldsymbol{x}_{i\cdot}^{(\ell)} &:= (x_{i1}^{(\ell)}, \cdots, x_{iK_i}^{(\ell)})^{\top} \cdot
\end{aligned}$$

Here, $\mathbf{0}_b$ is a $b$-vector of zeros, where $b$ denotes the number of covariates. Under a set of regularity conditions given by Liang and Zeger (1986), $\boldsymbol{\beta}_N$ is a consistent predictor of $\boldsymbol{B}$. The covariance structure $\boldsymbol{\Sigma}_i$ within $\boldsymbol{g}_i(\boldsymbol{\beta})$ does not affect the consistency but only the efficiency (Diggle et al., 2002; Liang and Zeger, 1986). The estimating equation (3) would need to be changed to accommodate random slopes and/or complex correlation structures. The empirical likelihood approach proposed would still be valid in this case.

In §4 and 5, $\boldsymbol{\beta}_N$ is treated as the parameter of interest. Hence a design-based inference for $\boldsymbol{\beta}_N$ is considerd, where the sampling distribution is only specified by the sampling design. Under this framework, $y_{ij}$ and $\boldsymbol{x}_{ij}$ are treated as fixed, non-random constant vectors; that is, the sampling distribution is conditional on $y_{ij}$ and $\boldsymbol{x}_{ij}$. In §6, a model-design-based inference is used, where $y_{ij}$ and $\boldsymbol{x}_{ij}$ are random and $\boldsymbol{B}$ is the parameter of interest.

## 4. Sample weighted generalised estimating equations

Skinner and Vieira (2007) proposed an estimator $\widehat{\boldsymbol{\beta}}$ defined as the solution to the sample *weighted generalised estimating equation* (WGEE)

$$\sum_{i\in S} \pi_i^{-1}\, \widehat{\boldsymbol{g}}_i(\boldsymbol{\beta}) = \mathbf{0}_b, \tag{4}$$

where

$$\widehat{\boldsymbol{g}}_i(\boldsymbol{\beta}) := \widehat{\boldsymbol{X}}_i^\top\, \widehat{\boldsymbol{\Sigma}}_i^{-1} \big(\widehat{\boldsymbol{y}}_i - \widehat{\boldsymbol{X}}_i\boldsymbol{\beta}\big)\cdot \tag{5}$$

The quantities $\widehat{\boldsymbol{y}}_i$ and $\widehat{\boldsymbol{X}}_i$ are the sub-matrices of $\boldsymbol{y}_i$ and $\boldsymbol{X}_i$, which contains the observations of the sample $S_i$. Consider the following estimator of $\boldsymbol{\Sigma}_i^{-1}$ (Rao and Molina, 2015).

$$\widehat{\boldsymbol{\Sigma}}_i^{-1} \;:=\; \widehat{\sigma}_e^{-2} \Big\{ \mathrm{diag}(w_{j|i} \,:\, j\in S_i) - \widehat{\gamma}_i\, w_{i\cdot}^{-1}\, (\widehat{\boldsymbol{w}}_i\, \widehat{\boldsymbol{w}}_i^\top) \Big\},$$

with

$$\begin{aligned}
w_{j|i} &:= (\pi_{j|i}\, a_i)^{-1},\\
\widehat{\boldsymbol{w}}_i &:= \mathrm{vector}(w_{j|i} \,:\, j\in S_i),\\
\widehat{\gamma}_i &:= \widehat{\sigma}_u^2 \big(\widehat{\sigma}_u^2 + \widehat{\sigma}_e^2 w_{i\cdot}^{-1}\big)^{-1},\\
w_{i\cdot} &:= \sum_{j\in S_i} w_{j|i},
\end{aligned} \tag{6}$$

where $\widehat{\sigma}_e^2$ and $\widehat{\sigma}_u^2$ are sample based estimators of $\sigma_e^2$ and $\sigma_u^2$. These estimators are given in Appendix A in the supplement. Here, $a_i$ are Potthoff, Woodbury and Manton's (1992) *scaling factors*,

$$a_i = \sum_{j\in S_i} \pi_{j|i}^{-2} \Bigg( \sum_{j\in S_i} \pi_{j|i}^{-1} \Bigg)^{-1}, \tag{7}$$

which are called "*scaling factors 1*" by Pfeffermann et al. (1998).

The design-consistency of $\widehat{\boldsymbol{\beta}}$ can be established by using a Taylor expansion of (4) and assuming the following key assumption (e.g. Godambe and Thompson, 2009; Berger, 2018a).

$$N^{-1} \sum_{i\in S} \pi_i^{-1}\, \widehat{\boldsymbol{g}}_i(\boldsymbol{\beta}_N) = O_{\mathcal{P}}(n^{-\frac{1}{2}}), \tag{8}$$

where $O_{\mathcal{P}}(\cdot)$ denotes the order of convergence in probability with respect to the sampling design.

If $a_i = 1$ instead of (7) and if $\widehat{\gamma}_i$ is given by (6), $\widehat{\boldsymbol{\beta}}$ reduces to You and Rao (2002, p.435) estimator. When $a_i = 1$ and $\widehat{\gamma}_i = 0$, it reduces to the *weighted composite likelihood* estimator proposed by Rao et al. (2013), or the *probability weighted least squares* estimator (Pfeffermann et al., 1998). Under the model (2), $\widehat{\boldsymbol{\beta}}$ is also the WIGLS estimator (Skinner and Vieira, 2007, p.5).

Poor estimation of the variance components $\sigma_u^2$ and $\sigma_e^2$ may result in some loss in efficiency of $\widehat{\boldsymbol{\beta}}$. Pfeffermann et al. (1998, p.29) showed that $\widehat{\boldsymbol{\beta}}$ is design-consistent, as $n \to \infty$, when $\sigma_e^2$ and $\sigma_u^2$ are used within (6) instead of $\widehat{\sigma}_u^2$ and $\widehat{\sigma}_e^2$ given in Appendix A in the supplement. However, $\sigma_e^2$ and $\sigma_u^2$ are rarely known and the bias of $\widehat{\sigma}_u^2$ and $\widehat{\sigma}_e^2$ may be high if the cluster sample sizes $k_i$ are small. Scaling factors $a_i$ may reduce the bias of $\widehat{\sigma}_u^2$ and $\widehat{\sigma}_e^2$ when $k_i$ are small (Pfeffermann et al., 1998). Several scaling factors have been proposed in the literature (e.g. Asparouhov, 2006; Clogg and Eliason, 1987; Graubard and Korn, 1996; Longford, 1995; Pfeffermann et al., 1998; Potthoff et al., 1992). However, there is no theoretical evidence supporting which scaling factor is better. Asparouhov (2006) compared the empirical biases of $\widehat{\sigma}_u^2$ and $\widehat{\sigma}_e^2$ for different scaling factors. In Appendix A in the supplement, several estimators are compared empirically. Pfeffermann et al. (1998, p.29) showed that the consistency of $\widehat{\boldsymbol{\beta}}$ is achieved provided that the number of clusters $n$ is large and the scaling factors $a_i$ are independent from $y_{ij}$.

# 5. Methodology proposed

## 5.1. Empirical likelihood approach

Several empirical likelihood approaches can be found in the literature (Chaudhuri et al., 2008; Chen and Sitter, 1999; Chen and Kim, 2014; Kim, 2009; Owen, 1988, 2001; Wu and Rao, 2006). Berger's (2018a) approach is used, because it allows profiling and complex sampling. It will be shown how it can be extended for multilevel models. The PSU-*level empirical log-likelihood function* is given by

$$\ell(\boldsymbol{p}) := \sum_{i\in S} \log p_i, \tag{9}$$

where the $p_i$ are unknown empirical likelihood probabilities allocated to the PSUs $i \in s$ and $\boldsymbol{p}$ denotes the $n \times 1$ vector of $p_i$.

Let $\widehat{p}_i^*(\boldsymbol{\beta})$ maximize $\ell(\boldsymbol{p})$ subject to the constraints $p_i > 0$ and

$$n \sum_{i\in S} p_i\, \pi_i^{-1}\, \boldsymbol{c}_i^*(\boldsymbol{\beta}) = \boldsymbol{C}^*, \tag{10}$$

for a given $\boldsymbol{\beta}$, with

$$\begin{aligned}
\boldsymbol{c}_i^*(\boldsymbol{\beta}) &:= \{\boldsymbol{c}_i^\top, \widehat{\boldsymbol{g}}_i(\boldsymbol{\beta})^\top\}^\top \quad \text{and} \quad \boldsymbol{C}^* := (\boldsymbol{C}^\top, \boldsymbol{0}^\top)^\top, & (11)\\
\boldsymbol{c}_i &:= (\boldsymbol{z}_i^\top, \widehat{\mathbf{f}}_i^\top)^\top \quad \text{and} \quad \boldsymbol{C} := (\boldsymbol{n}_H^\top, \boldsymbol{0}^\top)^\top, & (12)\\
\boldsymbol{z}_i &:= (z_{i1}, \dots, z_{iH})^\top \quad \text{and} \quad \boldsymbol{n}_H := \textstyle\sum_{i\in U} \boldsymbol{z}_i = (n_1, \dots, n_H)^\top, & (13)\\
\widehat{\mathbf{f}}_i &:= \sum_{j\in S_i} \pi_{j|i}^{-1} \mathbf{f}_{ij},
\end{aligned}$$

where $z_{ih} = \pi_i$ for $i \in U_h$ and $z_{ih} = 0$ otherwise. Here, $\widehat{\boldsymbol{g}}_i(\boldsymbol{\beta})$ is defined by (5) and $\mathbf{f}_{ij} := \mathbf{f}(\boldsymbol{\varphi}_N, \boldsymbol{v}_{ij})$ defined in §2. Assume that the $\boldsymbol{C}^*$ is an inner point of the conical hull formed by $n \sum_{i\in S} p_i \pi_i^{-1} \boldsymbol{c}_i^*(\boldsymbol{\beta})$, with $\boldsymbol{\beta}$ within the parameter space. Hence, the set of $\widehat{p}_i^*(\boldsymbol{\beta})$ is unique.

Note that (10) implies the *stratification constraint*

$$\sum_{i\in S} \frac{p_i}{\pi_i} \boldsymbol{z}_i = \frac{\boldsymbol{n}_H}{n}, \tag{14}$$

which is not motivated by moment conditions. This constraint implies $\sum_{i\in S} p_i = 1$, since $\mathbf{1}_H^\top \boldsymbol{z}_i = \pi_i$ and $\mathbf{1}_H^\top \boldsymbol{n}_H = n$. Hence, (14) can be viewed as a generalisation for stratification of Owen's (1988) leading constraint $\sum_{i\in S} p_i = 1$. The main difference between this §'s approach and the standard empirical likelihood approach (Owen, 1988) is the $\pi_i^{-1}$ within the constraint (10) and the use of the stratification constraint (14) instead of $\sum_{i\in S} p_i = 1$. Constraints weighted by $\pi_i^{-1}$ can also be found in Berger and Torres (2012, 2014, 2016), Chen and Kim (2014), Oğuz-Alper and Berger (2016) and Berger (2018a). Note that (9) is a PSU-level function.

The maximum value of $\ell(\boldsymbol{p})$ under $p_i > 0$ and (10) is given by

$$\ell(\boldsymbol{\beta}) = \sum_{i\in S} \log \widehat{p}_i^*(\boldsymbol{\beta})\cdot \tag{15}$$

The function $\ell(\boldsymbol{\beta})$ is not a parametric likelihood, but it behaves like a likelihood. It takes into account the sampling design and the population-level information, because $\boldsymbol{z}_i$ and $\widehat{\mathbf{f}}_i$ are included within $\boldsymbol{c}_i^*(\boldsymbol{\beta})$.

In Berger and Torres (2016) and Oğuz-Alper and Berger (2016), the empirical log-likelihood function is parametrized with $m_i := np_i\pi_i^{-1}$. By replacing $p_i$ by $n^{-1}m_i\pi_i$ within (9) and (10), (15) reduces to Berger and Torres's (2016) empirical log-likelihood function plus a quantity which does not depend on $\boldsymbol{\beta}$ and $m_i$. This dual parametrization in term of $m_i$ is equivalent.

The *maximum empirical likelihood estimator* $\widehat{\boldsymbol{\beta}}_{EL}$ of $\boldsymbol{\beta}_N$ is the vector that maximizes expression (15). This estimator is also the solution to the following sample estimating equation (Berger and Torres, 2016)

$$\widehat{\boldsymbol{G}}(\boldsymbol{\beta}) := \sum_{i\in S} \widehat{m}_i \, \widehat{\boldsymbol{g}}_i(\boldsymbol{\beta}) = \mathbf{0}_b, \tag{16}$$

where

$$\widehat{m}_i := n \, \widehat{p}_i \, \pi_i^{-1},$$

are the *empirical likelihood weights*, with

$$\widehat{p}_i := n^{-1}\left(1 + \boldsymbol{\eta}^\top \boldsymbol{c}_i \pi_i^{-1}\right)^{-1}\cdot \tag{17}$$

The vector $\boldsymbol{\eta}$ is such that $\widehat{m}_i > 0$ and

$$\sum_{i\in S} \widehat{m}_i \, \boldsymbol{c}_i = \boldsymbol{C}, \tag{18}$$

holds. Expression (17) is obtained from (B.9) in the supplement, by using $\boldsymbol{c}_i$ instead of $\boldsymbol{c}_i^*(\boldsymbol{\beta})$ (see §B.2 in Appendix B of the supplement). A modified Newton-Raphson algorithm as in Chen, Sitter and Wu (2002) can be used to compute $\boldsymbol{\eta}$. It can be shown that $\widehat{\boldsymbol{\beta}}_{EL}$ is $\sqrt{n}$-design-consistent under (8).

The $\widehat{m}_i$ play the role of survey weights. They are always positive and calibrated because $\sum_{i\in S} \widehat{m}_i \widehat{\mathbf{f}}_i = \mathbf{0}$ and (1) holds. The $\widehat{p}_i$ are similar to the so-called *g-weights* in Särndal, Swensson and Wretman (1992, p.232) or *calibration factors* in Deville and Särndal (1992). The calibration property is the consequence of the maximisation of (15) and the fact that $\boldsymbol{\varphi}_N$ is constant. In survey sampling literature, calibration is viewed as a weighting procedure, rather than the consequence of a likelihood principle.

The quantities $\widehat{m}_i$ satisfy $\sum_{i\in S} \widehat{m}_i \boldsymbol{z}_i = \boldsymbol{n}_H$, which specifies that $n_h$ observations selected with unequal probabilities within each stratum. Without population-level information, $\boldsymbol{c}_i = \boldsymbol{z}_i$, $\boldsymbol{\eta} = \mathbf{0}$ and $\widehat{m}_i = \pi_i^{-1}$, the standard Horvitz and Thompson's (1952) weight. In this case, $\widehat{\boldsymbol{\beta}}_{EL}$ is the WGEE estimator $\widehat{\boldsymbol{\beta}}$; which is the solution to (4).

## 5.2. Tests and confidence intervals

Suppose that the parameter of interest $\boldsymbol{\theta}_N$ is a $p$-sub-parameter of $\boldsymbol{\beta}_N$; that is, $\boldsymbol{\beta}_N = (\boldsymbol{\theta}_N^\top, \boldsymbol{v}_N^\top)^\top$. In order to make inference about $\boldsymbol{\theta}_N$, the following profile empirical log-likelihood ratio function is used.

$$\widehat{r}(\boldsymbol{\theta}) = 2\left\{\ell(\widehat{\boldsymbol{\beta}}) - \max_{\boldsymbol{v}\in\boldsymbol{\Lambda}} \ell(\boldsymbol{\theta}, \boldsymbol{v})\right\},$$

where $\ell(\boldsymbol{\theta}, \boldsymbol{v}) = \ell(\boldsymbol{\beta})$ with $\boldsymbol{\beta} = (\boldsymbol{\theta}^\top, \boldsymbol{v}^\top)^\top$ and $\ell(\widehat{\boldsymbol{\beta}}) = \sum_{i\in s} \log \widehat{m}_i$, where the $\widehat{m}_i$ are defined by (17). The symbol $\boldsymbol{\Lambda}$ denotes the parameter space of $\boldsymbol{v}_N$. The quantity $\max_{\boldsymbol{v}\in\boldsymbol{\Lambda}} \ell(\boldsymbol{\theta}, \boldsymbol{v})$ is given by

$$\max_{\boldsymbol{v}\in\boldsymbol{\Lambda}} \ell(\boldsymbol{\theta}, \boldsymbol{v}) = -\sum_{i\in S} \log\left[\pi_i + \overset{*}{\boldsymbol{\eta}}{}^\top \; \boldsymbol{c}_i^*\{\boldsymbol{\theta}, \mathring{\boldsymbol{v}}(\boldsymbol{\theta})\}\right] + \sum_{i\in S} \log(\pi_i n^{-1}),$$

where $\boldsymbol{c}_i^*\{\boldsymbol{\theta}, \mathring{\boldsymbol{v}}(\boldsymbol{\theta})\}$ is given by (11), with $\boldsymbol{\beta} = \{\boldsymbol{\theta}^\top, \mathring{\boldsymbol{v}}(\boldsymbol{\theta})^\top\}^\top$. The vector $\overset{*}{\boldsymbol{\eta}}$ and $\mathring{\boldsymbol{v}}(\boldsymbol{\theta})$ are the solutions to

$$\boldsymbol{\Gamma}_1(\boldsymbol{\theta}, \boldsymbol{v}) \quad := \quad \sum_{i\in S}\left\{\pi_i + \overset{*}{\boldsymbol{\eta}}{}^\top \; \boldsymbol{c}_i^*(\boldsymbol{\beta})\right\}^{-1} \boldsymbol{c}_i^*(\boldsymbol{\beta}) - \boldsymbol{C}^* = \mathbf{0}, \tag{19}$$

$$\boldsymbol{\Gamma}_2\{\overset{*}{\boldsymbol{\eta}}, \mathring{\boldsymbol{v}}(\boldsymbol{\theta})\} \quad := \quad n \, \overset{*}{\boldsymbol{\eta}}{}^\top \sum_{i\in S} \widehat{p}_i^*(\boldsymbol{\beta})\pi_i^{-1} \frac{\partial \boldsymbol{c}_i^*(\boldsymbol{\beta})}{\partial \, \mathring{\boldsymbol{v}}(\boldsymbol{\theta})} = \mathbf{0}\cdot \tag{20}$$

These solutions can be found by using a modified Newton-Raphson algorithm (Oğuz-Alper and Berger, 2016, Appendix A).

Theorem 1 shows that $\hat{r}(\boldsymbol{\theta}_N)$ is asymptotically pivotal and can be used for testing hypotheses. Confidence intervals can be constructed based on (21), when $\boldsymbol{\theta}_N$ is scalar.

**Theorem 1.** *Under the regularity conditions* (8), (B.2)–(B.6) *and* (B.8)*, given in Appendix B of the supplement,*

$$\hat{r}(\boldsymbol{\theta}_N) \xrightarrow{d} \chi^2_{df=p}, \tag{21}$$

*in distribution with respect to the design, when $n/N \to 0$. The constant $p$ is the dimension of $\boldsymbol{\theta}_N$.*

In §6, this theorem is generalised to allow $n/N \not\to 0$. The proof of Theorem 1 can be found in Appendix B in the supplement. This proof is not based on the assumption that non-random quantities $\sigma_u^2$ and $\sigma_e^2$ are used within (6). However, condition (8) is required. In practice, the $k_i$ are usually moderate sizes. Theorem 1 holds in this case, because the $\boldsymbol{K}_i$ are assumed asymptotically bounded. This is an advantage over the weighted multilevel pseudo-likelihood approach (Asparouhov, 2006; Rabe-Hesketh and Skrondal, 2006) which is based on the assumption that the $k_i$ are large.

## 6. Large sampling fractions

In this §, Theorem 1 is generalised for large sampling fraction ($n/N \not\to 0$). In other words, the empirical log-likelihood ratio function still converges to a $\chi^2$-distribution, when the sampling fractions are large. It will not be necessary to adjust the empirical log-likelihood ratio function with eigenvalues as in Berger (2018a); Rao and Scott (1981) and Zhao and Wu (2018) or with design effect as in Wu and Rao (2006) or with finite population corrections as in Berger and Torres (2016). The simulation study in §7.2 supports out findings. The empirical likelihood point estimator is also model-design consistent.

In §4 and 5, $\boldsymbol{\beta}_N$ is the target parameter because the model-variance of $\boldsymbol{\beta}_N - \boldsymbol{B}$ is negligible when $n/N$ is small. However, with large sampling fractions this model-variance may not be negligible compared to the design-variance. By considering $\boldsymbol{B}$ as the target rather than $\boldsymbol{\beta}_N$, the empirical log-likelihood ratio function implicitly incorporates Hájek's (1964) approximation of the joint inclusion probabilities and a model variance.

Assume that the regularity conditions (B.2)–(B.6) in the supplement, hold after replacing $\boldsymbol{\beta}_N$ by $\boldsymbol{B}$. Let $\boldsymbol{B} = (\boldsymbol{\vartheta}^\top, \boldsymbol{\upsilon}^\top)^\top$, where the $p$-vector $\boldsymbol{\vartheta}$ is the parameter of interest and $\boldsymbol{\upsilon}$ the remaining parameters. Let $\hat{r}(\boldsymbol{\vartheta}) = 2\{\ell(\hat{\boldsymbol{\beta}}) - \max_{\boldsymbol{\upsilon}\in\boldsymbol{\Lambda}} \ell(\boldsymbol{\vartheta}, \boldsymbol{\upsilon})\}$. The proof in Appendix B can be used to show that

$$\hat{r}(\boldsymbol{\vartheta}) = \frac{1}{N^2}\hat{\boldsymbol{G}}_{reg}(\boldsymbol{B})^\top(\mathbf{I} - \hat{\boldsymbol{A}}^{\bullet})\tilde{\boldsymbol{V}}^{-1}\hat{\boldsymbol{G}}_{reg}(\boldsymbol{B}) + O_{\mathcal{P}}(n^{-\frac{1}{2}}),$$

where $\hat{\boldsymbol{G}}_{reg}(\boldsymbol{B})$ is regression-type estimator given by

$$\hat{\boldsymbol{G}}_{reg}(\boldsymbol{B}) := \hat{\boldsymbol{G}}_\pi(\boldsymbol{B}) - \hat{\boldsymbol{\Psi}}(\boldsymbol{B})^\top\,\hat{\mathbf{f}}_\pi, \tag{22}$$

with

$$\begin{aligned}
\hat{\boldsymbol{G}}_\pi(\boldsymbol{B}) &:= \sum_{i\in S}\frac{\hat{\boldsymbol{g}}_i(\boldsymbol{B})}{\pi_i}, \qquad \hat{\mathbf{f}}_\pi := \sum_{i\in S}\frac{\hat{\mathbf{f}}_i}{\pi_i}, \\
\hat{\boldsymbol{\Psi}}(\boldsymbol{B}) &:= \widehat{\mathbf{var}}(\hat{\mathbf{f}}_\pi)^{-1}\widehat{\mathbf{cov}}\{\hat{\mathbf{f}}_\pi, \hat{\boldsymbol{G}}_\pi(\boldsymbol{B})\}, \\
\widehat{\mathbf{var}}(\hat{\mathbf{f}}_\pi) &:= \sum_{i\in S}\frac{\hat{\mathbf{f}}_i\,\hat{\mathbf{f}}_i^\top}{\pi_i^2} - \sum_{i\in S}\frac{\hat{\mathbf{f}}_i\,\boldsymbol{z}_i^\top}{\pi_i^2}\left(\sum_{i\in S}\frac{\boldsymbol{z}_i\,\boldsymbol{z}_i^\top}{\pi_i^2}\right)^{-1}\sum_{i\in S}\frac{\boldsymbol{z}_i\hat{\mathbf{f}}_i^\top}{\pi_i^2}, \\
\widehat{\mathbf{cov}}\{\hat{\mathbf{f}}_\pi, \hat{\boldsymbol{G}}_\pi(\boldsymbol{B})\} &:= \sum_{i\in S}\frac{\hat{\mathbf{f}}_i\,\hat{\boldsymbol{g}}_i(\boldsymbol{B})^\top}{\pi_i^2} - \sum_{i\in S}\frac{\hat{\mathbf{f}}_i\,\boldsymbol{z}_i^\top}{\pi_i^2}\left(\sum_{i\in S}\frac{\boldsymbol{z}_i\,\boldsymbol{z}_i^\top}{\pi_i^2}\right)^{-1}\sum_{i\in S}\frac{\boldsymbol{z}_i\,\hat{\boldsymbol{g}}_i(\boldsymbol{B})^\top}{\pi_i^2}.
\end{aligned}$$

The matrix $\widehat{\boldsymbol{A}}^{\bullet}$ is symmetric, idempotent and defined by

$$\widehat{\boldsymbol{A}}^{\bullet} := \widetilde{\boldsymbol{V}}^{-\frac{1}{2}}\widehat{\boldsymbol{\nabla}}^{\bullet}_{g}\left(\widehat{\boldsymbol{\nabla}}^{\bullet\top}_{g}\widetilde{\boldsymbol{V}}^{-1}\widehat{\boldsymbol{\nabla}}^{\bullet}_{g}\right)^{-1}\widehat{\boldsymbol{\nabla}}^{\bullet\top}_{g}\widetilde{\boldsymbol{V}}^{-\frac{1}{2}}, \tag{23}$$

where

$$\begin{aligned}
\widehat{\boldsymbol{\nabla}}^{\bullet}_{g} &:= \sum_{i\in S}\pi_i^{-1}\frac{\partial\,\widehat{\boldsymbol{g}}^{\circ}_{i}(\boldsymbol{B})}{\partial\boldsymbol{\upsilon}} \quad \text{and} \quad \widehat{\boldsymbol{g}}^{\circ}_{i}(\boldsymbol{B}) := \widehat{\boldsymbol{g}}_{i}(\boldsymbol{B}) - \widehat{\boldsymbol{\Psi}}(\boldsymbol{B})^{\top}\,\widehat{\mathbf{f}}_{i},\\
\widetilde{\boldsymbol{V}} &:= \frac{1}{N^2}\sum_{i\in S}\frac{1}{\pi_i^2}\widehat{\boldsymbol{g}}^{\circ}_{i}(\boldsymbol{B})\,\widehat{\boldsymbol{g}}^{\circ}_{i}(\boldsymbol{B})^{\top} - \frac{1}{N^2}\widehat{\boldsymbol{G}}^{\circ}(\boldsymbol{B})^{\top}\boldsymbol{Z}^{-1}\widehat{\boldsymbol{G}}^{\circ}(\boldsymbol{B}),\\
\widehat{\boldsymbol{G}}^{\circ}(\boldsymbol{B}) &:= \sum_{i\in S}\frac{1}{\pi_i^2}\boldsymbol{z}_i\,\widehat{\boldsymbol{g}}^{\circ}_{i}(\boldsymbol{B})^{\top} \quad \text{and} \quad \boldsymbol{Z} := \sum_{i\in S}\frac{1}{\pi_i^2}\boldsymbol{z}_i\,\boldsymbol{z}_i^{\top} = \operatorname{diag}(\boldsymbol{n}_H)\cdot
\end{aligned}$$

The vectors $\boldsymbol{z}_i$ and $\boldsymbol{n}_H$ are defined by (13).

The matrix $\widetilde{\boldsymbol{V}}$ is the usual Hansen and Hurwitz's (1943) stratified variance estimator, considered under negligible sampling fraction. In the rest of this §, it is shown that $\widetilde{\boldsymbol{V}}$ is also a consistent variance estimator of the model-design variance of $N^{-1}\widehat{\boldsymbol{G}}_{reg}(\boldsymbol{B})$, given by

$$\boldsymbol{V} := \mathbb{E}_m\mathbb{V}_d\left\{N^{-1}\widehat{\boldsymbol{G}}_{reg}(\boldsymbol{B})\right\} + \mathbb{V}_m\mathbb{E}_d\left\{N^{-1}\widehat{\boldsymbol{G}}_{reg}(\boldsymbol{B})\right\}, \tag{24}$$

where $\mathbb{E}_d$ and $\mathbb{V}_d$ are the expectation and variance with respect to the two-stage design. The operators $\mathbb{E}_m$ and $\mathbb{V}_m$ represent the model expectation and variance under (2). The consistency of $\widetilde{\boldsymbol{V}}$ may seem counter-intuitive, but can be explained by the fact that $\widetilde{\boldsymbol{V}}$ over-estimates the design variance $\mathbb{V}_d\{N^{-1}\widehat{\boldsymbol{G}}_{reg}(\boldsymbol{B})\}$ by an amount which estimates the second term of (24), as shown in the following lemma.

**Lemma 1.** *Under*

$$\frac{1}{N}\left\|\sum_{i\in s}\frac{1}{\pi_i}\boldsymbol{c}^{*}_{i}(\boldsymbol{B}) - \boldsymbol{C}^{*}\right\| = O_{\mathcal{P}}(n^{-\frac{1}{2}}), \tag{25}$$

*the matrix $\widetilde{\boldsymbol{V}}$ can be decomposed as*

$$\widetilde{\boldsymbol{V}} = \widehat{\boldsymbol{V}}_d + \widehat{\boldsymbol{V}}_m + O_{\mathcal{P}}(n^{-2}), \tag{26}$$

*where*

$$\begin{aligned}
\widehat{\boldsymbol{V}}_d &:= \frac{1}{N^2}\sum_{i\in S}\frac{1-\pi_i}{\pi_i^2}\widehat{\boldsymbol{g}}^{\circ}_{i}(\boldsymbol{B})\,\widehat{\boldsymbol{g}}^{\circ}_{i}(\boldsymbol{B})^{\top} - \frac{1}{N^2}\widehat{\boldsymbol{G}}^{\circ}_{c}(\boldsymbol{B})^{\top}\widehat{\boldsymbol{Z}}_c^{-1}\widehat{\boldsymbol{G}}^{\circ}_{c}(\boldsymbol{B}), \qquad (27)\\
\widehat{\boldsymbol{V}}_m &:= \frac{1}{N^2}\sum_{i\in S}\frac{1}{\pi_i}\widehat{\boldsymbol{g}}^{\circ}_{i}(\boldsymbol{B})\,\widehat{\boldsymbol{g}}^{\circ}_{i}(\boldsymbol{B})^{\top},\\
\widehat{\boldsymbol{G}}^{\circ}_{c}(\boldsymbol{B}) &:= \sum_{i\in S}\frac{1-\pi_i}{\pi_i^2}\boldsymbol{z}_i\,\widehat{\boldsymbol{g}}^{\circ}_{i}(\boldsymbol{B})^{\top} \quad \textit{and} \quad \widehat{\boldsymbol{Z}}_c := \sum_{i\in S}\frac{1-\pi_i}{\pi_i^2}\boldsymbol{z}_i\,\boldsymbol{z}_i^{\top}\cdot
\end{aligned}$$

*Furthermore,*

$$\frac{1}{N^2}\widehat{\boldsymbol{G}}^{\circ}_{c}(\boldsymbol{B})^{\top}\widehat{\boldsymbol{Z}}_c^{-1}\widehat{\boldsymbol{G}}^{\circ}_{c}(\boldsymbol{B}) = O_{\mathcal{P}}(n^{-2})\cdot \tag{28}$$

The proof of Lemma 1 can be found in Appendix B of the supplement. The condition (25) is the law of large numbers which can be justified by using the Isaki and Fuller's (1982) sufficient conditions. Note that conditions (B.2) and (B.6) in the supplement imply $\widehat{\boldsymbol{V}}_d = O_{\mathcal{P}}(n^{-1})$ and $\widehat{\boldsymbol{V}}_m = O_{\mathcal{P}}(N^{-1})$; that is, $\widehat{\boldsymbol{V}}_m$ is not asymptotically negligible when $n/N$ is large.

The matrix (27) is Hájek's (1981) stratified design-based estimator which implicitly includes Hájek's (1964) approximation of the joint inclusion probabilities which are often considered in the literature (Aires, 2000; Berger, 1998;

Deville, 1999; Fuller, 2009; Haziza, Mecatti and Rao, 2004; Matei and Tillé, 2005). The quantities $1-\pi_i$ are PSU-level finite population corrections, which reduces to $1-n/N$ under equal probability sampling of PSUs.

The key results is to show that $\widetilde{\boldsymbol{V}}$ is indeed an unbiased estimator of $\boldsymbol{V}$. Following Hájek's (1964) asymptotic framework, based on $\sum_{i\in U}\pi_i(1-\pi_i)\to\infty$, expression (28) implies that the first two-stage expectation is given by

$$\mathbb{E}_d(\widehat{\boldsymbol{V}}_d)=\mathbb{V}_d\{N^{-1}\widehat{\boldsymbol{G}}_{reg}(\boldsymbol{B})\}-\frac{1}{N^2}\sum_{i\in U}\mathbb{V}_d\{\widehat{\boldsymbol{g}}_i^{\circ}(\boldsymbol{B})\}+o_{\mathcal{P}}(1),$$

where $o_{\mathcal{P}}(\cdot)$ is the order of convergence in probability with respect to the sampling design $\mathcal{P}(s)$ (e.g. Isaki and Fuller, 1982). Hence

$$\mathbb{E}_m\mathbb{E}_d(\widehat{\boldsymbol{V}}_d)=\mathbb{E}_m\mathbb{V}_d\big\{N^{-1}\widehat{\boldsymbol{G}}_{reg}(\boldsymbol{B})\big\}-\frac{1}{N^2}\sum_{i\in U}\mathbb{E}_m\mathbb{V}_d\{\widehat{\boldsymbol{g}}_i^{\circ}(\boldsymbol{B})\}+o(1)\cdot \tag{29}$$

Now,

$$\begin{aligned}\mathbb{E}_d(\widehat{\boldsymbol{V}}_m) &= \frac{1}{N^2}\sum_{i\in U}\mathbb{E}_d\big\{\widehat{\boldsymbol{g}}_i^{\circ}(\boldsymbol{B})\,\widehat{\boldsymbol{g}}_i^{\circ}(\boldsymbol{B})^{\top}\big\}\\ &= \frac{1}{N^2}\sum_{i\in U}\mathbb{V}_d\{\widehat{\boldsymbol{g}}_i^{\circ}(\boldsymbol{B})\}+\frac{1}{N^2}\sum_{i\in U}\boldsymbol{g}_i^{\circ}(\boldsymbol{B})\,\boldsymbol{g}_i^{\circ}(\boldsymbol{B})^{\top},\end{aligned}$$

where $\boldsymbol{g}_i^{\circ}(\boldsymbol{B}):=\mathbb{E}_d\{\widehat{\boldsymbol{g}}_i^{\circ}(\boldsymbol{B})\}$. Hence,

$$\mathbb{E}_m\mathbb{E}_d(\widehat{\boldsymbol{V}}_m)=\frac{1}{N^2}\sum_{i\in U}\mathbb{E}_m\mathbb{V}_d\{\widehat{\boldsymbol{g}}_i^{\circ}(\boldsymbol{B})\}+\frac{1}{N^2}\sum_{i\in U}\mathbb{E}_m\big\{\boldsymbol{g}_i^{\circ}(\boldsymbol{B})\,\boldsymbol{g}_i^{\circ}(\boldsymbol{B})^{\top}\big\}\cdot \tag{30}$$

Combining (26), (29) and (30) gives

$$\mathbb{E}_m\mathbb{E}_d(\widetilde{\boldsymbol{V}})=\mathbb{E}_m\mathbb{V}_d\big\{N^{-1}\widehat{\boldsymbol{G}}_{reg}(\boldsymbol{B})\big\}+\frac{1}{N^2}\sum_{i\in U}\mathbb{E}_m\big\{\boldsymbol{g}_i^{\circ}(\boldsymbol{B})\,\boldsymbol{g}_i^{\circ}(\boldsymbol{B})^{\top}\big\}+o(1)\cdot \tag{31}$$

Since $\mathbb{E}_d\{N^{-1}\widehat{\boldsymbol{G}}_{reg}(\boldsymbol{B})\}=N^{-1}\sum_{i\in U}\boldsymbol{g}_i^{\circ}(\boldsymbol{B})$ and $\mathbb{E}_m\{\boldsymbol{g}_i^{\circ}(\boldsymbol{B})\}=o(1)$, it can be shown that

$$\mathbb{V}_m\mathbb{E}_d\{N^{-1}\widehat{\boldsymbol{G}}_{reg}(\boldsymbol{B})\}=\frac{1}{N^2}\sum_{i\in U}\mathbb{E}_m\big\{\boldsymbol{g}_i^{\circ}(\boldsymbol{B})\,\boldsymbol{g}_i^{\circ}(\boldsymbol{B})^{\top}\big\}+o(1)\cdot \tag{32}$$

Now, (24), (31) and (32) imply

$$\begin{aligned}\mathbb{E}_m\mathbb{E}_d(\widetilde{\boldsymbol{V}}) &= \mathbb{E}_m\mathbb{V}_d\big\{N^{-1}\widehat{\boldsymbol{G}}_{reg}(\boldsymbol{B})\big\}+\mathbb{V}_m\mathbb{E}_d\{N^{-1}\widehat{\boldsymbol{G}}_{reg}(\boldsymbol{B})\}+o(1)\\ &= \boldsymbol{V}+o(1)\cdot\end{aligned}$$

Consequently, $n\|\widetilde{\boldsymbol{V}}-\boldsymbol{V}\|=o(1)$, because $\widetilde{\boldsymbol{V}}$ converges to a matrix of constants (e.g. Berger, 2018d). Thus, by assuming that $\boldsymbol{V}^{-\frac{1}{2}}\,N^{-1}\widehat{\boldsymbol{G}}_{reg}(\boldsymbol{B})\xrightarrow{d}\mathcal{N}(\boldsymbol{0},\mathbf{I})$, it implies that

$$\widehat{r}(\boldsymbol{\vartheta})\xrightarrow{d}\chi^2_{df=p}, \tag{33}$$

in distribution with respect to the model and design. The constant $p$ is the trace of the idempotent matrix (23); that is, the dimension of $\boldsymbol{\vartheta}$. The pivotal property (33) holds, when $\boldsymbol{\vartheta}$ is the target parameter, even if $n/N$ is large. When the finite population parameter $\boldsymbol{\theta}_N$ is the target, it is necessary to assume that $n/N$ is negligible for the pivotal property (21) to hold. When $n/N\to 0$, the model variance of $\boldsymbol{\theta}_N-\boldsymbol{\vartheta}$ is negligible, $\boldsymbol{\theta}_N=\boldsymbol{\vartheta}+o(1)$ and (33) reduces to (21).

## 7. Simulation study

Consider a finite population generated from

$$y_{ij}=B_0+B_1x_{ij}^{(1)}+B_2x_{ij}^{(2)}+u_i+e_{ij}, \tag{34}$$

where $B_0 = 20$, $B_1 = B_2 = 1$, $x_{ij}^{(1)} \sim \Gamma(\text{shape} = 2, \text{scale} = \alpha_{1i})$, $x_{ij}^{(2)} \sim \Gamma(\text{shape} = 2, \text{scale} = \alpha_{2i})$, $u_i \sim N(0, \text{sd} = \sigma_u)$ and $e_{ij} \sim \Gamma(\text{shape} = \sigma_e^2/4, \text{scale} = 2) - \sigma_e^2/2$, with $\sigma_e^2 = 12 - \sigma_u^2$. The quantities $\alpha_{1i}$ and $\alpha_{2i}$ are selected randomly with-replacement among the values $\{1,2,3\}$ and $\{1,2,3,4\}$, respectively. The number of clusters is $N = 3000$. The cluster sizes $K_i$ are generated randomly from $K_i = \lfloor 100 \exp(\tau_i) \rfloor$, with $\tau_i \sim N(0, \text{sd} = 0.2)$, which gives $K_i$ ranging between 47 and 207, with $\sum_{i \in U} K_i = 305\,305$. The values of $\sigma_u^2$ are chosen to obtain different intra-cluster correlations, given by $\rho := \sigma_u^2/(\sigma_e^2 + \sigma_u^2) = \sigma_u^2/12$. The correlations considered range from 0.04 to 0.83 (see Table 1). Known population-level parameters are not considered. The parameter, $\boldsymbol{\beta}_N = (\beta_{0N}, \beta_{1N}, \beta_{2N})^\top$ obtained from (3) are given in Table 1.

**Table 1**
Values of $\sigma_u^2$, $\rho$, $b_0$ and $b_1$ used to generate the population data. Values of the components of the finite population parameter $\boldsymbol{\beta}_N$. $N = 3000$.

| $\sigma_u^2$ | $\rho$ | $b_0$ | $b_1$ | corr$(\pi_i, u_i)$ | $\beta_{0N}$ | $\beta_{1N}$ | $\beta_{2N}$ |
|---|---|---|---|---|---|---|---|
| 0.50 | 0.04 | 4 | 0.40 | 0.87 | 20.00 | 1.00 | 1.00 |
| 3.00 | 0.25 | 7 | 1.00 | 0.86 | 20.05 | 1.00 | 1.00 |
| 6.00 | 0.50 | 11 | 1.55 | 0.84 | 20.06 | 1.00 | 1.00 |
| 10.00 | 0.83 | 15 | 2.00 | 0.84 | 20.06 | 1.00 | 1.00 |

Consider three estimators: the standard restricted maximum likelihood (RML) estimator $\widehat{\boldsymbol{\beta}}^{rml} = (\widehat{\beta}_0^{rml}, \widehat{\beta}_1^{rml}, \widehat{\beta}_2^{rml})^\top$, the weighted composite likelihood (CL) estimator $\widehat{\boldsymbol{\beta}}^{cl} = (\widehat{\beta}_0^{cl}, \widehat{\beta}_1^{cl}, \widehat{\beta}_2^{cl})^\top$ and the empirical likelihood (EL) estimator proposed $\widehat{\boldsymbol{\beta}}^{el} = (\widehat{\beta}_0^{el}, \widehat{\beta}_1^{el}, \widehat{\beta}_2^{el})^\top$; which is the solution to (16). The empirical likelihood and the WGEE approaches give the same point estimator as EL, because $\widehat{m}_i = \pi_i^{-1}$ when known population-level parameters are not considered.

Let $\widehat{\theta}$ be an estimator of a parameter $\theta_N \in \boldsymbol{\beta}_N$. Let $\widehat{\theta}_m$ denote an estimate based on the $m$-th sample, where $m = 1, \dots, M$ and $M = 10\,000$. The empirical relative bias (%) of $\widehat{\theta}$ is RB% $:= [\{E(\widehat{\theta}) - \theta\}/\theta] \times 100\%$, where $E(\widehat{\theta}) := M^{-1} \sum_{m=1}^{M} \widehat{\theta}_m$ is the empirical design-based expectation.

The simulation was done in R (R Core Team, 2019). Some of the R codes used in this §, are available on the second author's web-page: `http://www.yvesberger.co.uk`.

## 7.1. Inclusion probability correlated with the random effect

The number of two-stage samples selected is 10 000. The first stage is a randomized systematic sample of $n = 150$ PSUs selected with unequal probabilities proportional to $\delta_i = b_0 + u_i + b_1 \epsilon_i$, where $\epsilon_i \sim \exp(rate = 1) - 1$. The values $b_0$ and $b_1$ are given in Table 1. They are chosen so that $0.84 \leqslant \text{corr}(\pi_i, u_i) \leqslant 0.87$.

For the second stage, simple random samples of $k_i = \alpha K_i$ SSUs were selected within the PSU $i$ selected, where $\alpha = 0.25$ in Tables 2 and 3; and $\alpha = 0.1$, 0.25 and 0.4, in Tables 5 and 6. With $\alpha = 0.1$, $5 \leqslant k_i \leqslant 21$, $12 \leqslant k_i \leqslant 52$ with $\alpha = 0.25$ and $19 \leqslant k_i \leqslant 83$ with $\alpha = 0.4$.

The RB of the regression parameters are given in Table 2. Note that the selection probabilities of the PSUs are proportional to the random effect $u_i$ which characterises the intercept. This is the reason of the observed differences between the relative biases of the intercepts. No noticeable differences are observed between the bias of the RML and empirical likelihood estimators of $\beta_{1N}$ and $\beta_{2N}$, because the inclusion probabilities only affects the intercept. The RML point estimator $\widehat{\beta}_0^{rml}$ is slightly biased, because the unequal probabilities. The RB increases with the intra-cluster correlation. The EL estimators has the smallest bias.

The Monte Carlo design-based performance of the EL confidence interval based upon (21) is compared with the alternative confidence intervals. The nominal level is 95%. The WGEE confidence interval is based on inverse testing (Binder and Patak, 1994) and Hartley and Rao's (1962) variance estimator. Composite likelihood confidence intervals rely on conditional linearised variance estimator proposed by Rao et al. (2013, p.270). Restricted maximum likelihood estimation confidence intervals are those obtained from the `lme()` function in R (R Core Team, 2019).

The observed coverages are given in Table 3. The p-values of D'Agostino's (1970) K-squared test of normality, show that the point estimators are mostly not normally distributed. Coverages of the RML confidence intervals for the intercept are significantly different from the nominal level, which is 95%, in all cases. The coverages decrease with the intra-cluster correlation. Poor coverages are due to the bias of $\widehat{\beta}_0^{rml}$. In this case, the variance of $\widehat{\beta}_0^{rml}$ is also

**Table 2**
Relative biases (%), biases and standard errors for different intra-cluster correlations, $\rho$. $N = 3000$. $n = 150$.

| | | Relative biases (%) | | | Biases ×1000 | | | Standard errors ×100 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $\rho$ | | EL | RML | CL | EL | RML | CL | EL | RML | CL |
| 0.04 | $\beta_{0N}$ | 0.03 | 0.63 | 0.05 | 6.2 | 126.0 | 9.5 | 12.6 | 12.2 | 13.0 |
| | $\beta_{1N}$ | 0.00 | 0.01 | 0.07 | 0.0 | 0.1 | 0.7 | 1.8 | 1.7 | 1.9 |
| | $\beta_{2N}$ | −0.06 | −0.07 | −0.10 | −0.6 | −0.7 | −1.0 | 1.4 | 1.3 | 1.5 |
| 0.25 | $\beta_{0N}$ | 0.02 | 2.05 | 0.03 | 3.5 | 410.8 | 6.9 | 18.6 | 16.5 | 21.8 |
| | $\beta_{1N}$ | 0.03 | 0.03 | 0.50 | 0.3 | 0.3 | 5.0 | 1.7 | 1.6 | 2.7 |
| | $\beta_{2N}$ | −0.03 | 0.00 | −0.13 | −0.3 | 0.0 | −1.3 | 1.3 | 1.3 | 2.2 |
| 0.50 | $\beta_{0N}$ | 0.01 | 2.62 | 0.07 | 2.8 | 525.8 | 14.4 | 22.6 | 20.4 | 28.1 |
| | $\beta_{1N}$ | 0.03 | 0.00 | 0.55 | 0.3 | 0.0 | 5.5 | 1.4 | 1.3 | 3.4 |
| | $\beta_{2N}$ | −0.03 | −0.01 | −0.19 | −0.3 | −0.1 | −1.9 | 1.1 | 1.0 | 2.7 |
| 0.83 | $\beta_{0N}$ | 0.03 | 3.22 | 0.10 | 6.3 | 646.2 | 20.6 | 27.3 | 24.7 | 34.9 |
| | $\beta_{1N}$ | 0.01 | 0.02 | 0.74 | 0.1 | 0.2 | 7.4 | 0.8 | 0.8 | 4.0 |
| | $\beta_{2N}$ | −0.01 | −0.03 | −0.26 | −0.1 | −0.3 | −2.6 | 0.6 | 0.6 | 3.2 |

substantially underestimated. The EL confidence intervals have better coverages than the WGEE and CL confidence intervals.

**Table 3**
Observed coverages (%) of 95% confidence intervals, for different intra-cluster correlations, $\rho$. D'Agostino's K-squared p-values within parentheses. $N = 3000$. $n = 150$.

| $\rho$ | | EL | RML | WGEE | CL |
|---|---|---|---|---|---|
| 0.04 | $\beta_{0N}$ | 94.9 (0.89) | 82.9$^{\dagger}$ (0.60) | 94.3$^{\dagger}$ (0.89) | 94.6 (0.90) |
| | $\beta_{1N}$ | 94.8 (0.00) | 95.0 (0.00) | 94.2$^{\dagger}$ (0.00) | 94.8 (0.00) |
| | $\beta_{2N}$ | 94.9 (0.71) | 94.9 (0.56) | 94.2$^{\dagger}$ (0.71) | 94.5$^{\dagger}$ (0.93) |
| 0.25 | $\beta_{0N}$ | 94.6 (0.00) | 31.4$^{\dagger}$ (0.33) | 94.1$^{\dagger}$ (0.00) | 94.2$^{\dagger}$ (0.02) |
| | $\beta_{1N}$ | 94.8 (0.79) | 94.9 (0.56) | 94.3$^{\dagger}$ (0.79) | 93.8$^{\dagger}$ (0.00) |
| | $\beta_{2N}$ | 94.1$^{\dagger}$ (0.01) | 94.7 (0.01) | 93.7$^{\dagger}$ (0.01) | 94.2$^{\dagger}$ (0.03) |
| 0.50 | $\beta_{0N}$ | 95.1 (0.00) | 28.7$^{\dagger}$ (0.13) | 94.6 (0.00) | 94.7 (0.01) |
| | $\beta_{1N}$ | 94.6 (0.05) | 95.1 (0.18) | 94.0$^{\dagger}$ (0.05) | 93.9$^{\dagger}$ (0.84) |
| | $\beta_{2N}$ | 94.6 (0.00) | 95.4 (0.00) | 94.1$^{\dagger}$ (0.00) | 94.2$^{\dagger}$ (0.00) |
| 0.83 | $\beta_{0N}$ | 94.9 (0.11) | 27.7$^{\dagger}$ (0.63) | 94.3$^{\dagger}$ (0.11) | 94.5$^{\dagger}$ (0.09) |
| | $\beta_{1N}$ | 94.7 (0.03) | 95.4 (0.11) | 94.1$^{\dagger}$ (0.03) | 94.3$^{\dagger}$ (0.63) |
| | $\beta_{2N}$ | 94.5$^{\dagger}$ (0.01) | 94.8 (0.00) | 93.9$^{\dagger}$ (0.01) | 94.5$^{\dagger}$ (0.05) |

$^{\dagger}$ Coverages significantly different from 95%. p-value $\leq 0.05$.

## 7.2. Large sampling fractions

The observed coverages for sampling fractions ranging from 5% to 70% and with a sample size of $n = 150$ PSUs are given in Table 4. The number of two-stage samples selected is 1000. The data follow the same model (34), with the variables generated as before. The only difference is that $e_{ij} \sim N(0, \text{sd} = \sigma_e)$ rather than following a Gamma distribution, in order to eliminate the effect of heteroscedasticity and isolate the effect of the sampling fraction. A model-design based approach is considered, with $\boldsymbol{B}$ being the parameter of interest, as in §6. As expected, the coverages observed are mostly not significantly different from 95%, because of (33). The columns' medians are reported in the last row. With $n/N = 5\%$, the median is 94.9%. With larger sampling fraction, the medians is slightly lower than 94.9%, without being significantly different from 95%. There is no other trend to report.

**Table 4**
Observed coverages (%) of 95% confidence intervals, for different intra-cluster correlations $\rho$ and different sampling fractions $n/N$. The columns' medians are given in the last row. $n = 150$.

| | | $n/N$ | | | | | |
|---|---|---|---|---|---|---|---|
| $\rho$ | | 5% | 10% | 20% | 40% | 50% | 70% |
| 0.04 | $B_0$ | 95.6 | 93.5$^\dagger$ | 94.8 | 94.9 | 95.1 | 94.1 |
| | $B_1$ | 94.2 | 94.0 | 94.6 | 94.6 | 95.1 | 94.8 |
| | $B_2$ | 95.3 | 94.8 | 93.6$^\dagger$ | 94.2 | 95.0 | 94.8 |
| 0.25 | $B_0$ | 95.2 | 94.3 | 94.4 | 94.3 | 94.2 | 95.0 |
| | $B_1$ | 95.1 | 93.5$^\dagger$ | 94.7 | 94.6 | 93.3$^\dagger$ | 93.1$^\dagger$ |
| | $B_2$ | 92.5$^\dagger$ | 93.9 | 92.9$^\dagger$ | 94.9 | 94.1 | 93.3$^\dagger$ |
| 0.50 | $B_0$ | 94.4 | 94.9 | 94.0 | 95.5 | 94.6 | 95.1 |
| | $B_1$ | 94.1 | 94.3 | 95.6 | 93.9 | 94.6 | 94.6 |
| | $B_2$ | 94.4 | 95.7 | 94.8 | 93.7 | 93.7 | 93.9 |
| 0.83 | $B_0$ | 95.0 | 94.3 | 94.1 | 94.6 | 94.3 | 94.8 |
| | $B_1$ | 94.8 | 95.0 | 93.5$^\dagger$ | 94.9 | 96.6$^\dagger$ | 94.1 |
| | $B_2$ | 95.4 | 94.4 | 93.1$^\dagger$ | 94.3 | 93.8 | 93.5$^\dagger$ |
| Medians: | | 94.9 | 94.3 | 94.3 | 94.6 | 94.5 | 94.4 |

$^\dagger$ Coverages significantly different from 95%. p-value $\leq 0.05$.

### 7.3. Population with outlying values

The finite population is generated from the model (34) with $e_{ij} \sim N(0, \text{sd} = \sigma_e)$. The intra-cluster correlation considered is $\rho = 0.50$. The number of clusters is $N = 3000$ and the $K_i$ and $k_i$ are generated as in §7. Let 10% of the $y_{ij}$ be replaced randomly by values generated randomly from $Y_{0.75} + 1.5 \times (Y_{0.75} - Y_{0.25}) + \tau_{ij}$, where $\tau_{ij} \sim \Gamma(shape = 2, scale = 2) - 4$. The quantities $Y_{0.25}$ and $Y_{0.75}$ are the lower and upper population quartiles of the $y_{ij}$. Another set of 10% of the $y_{ij}$ are replaced by values generated randomly from $\max\{y_{ij}\} + \tau_{ij}$. The number of PSUs selected is $n = 150$. Simple random sampling without replacement is used at both stages, in order to isolate the effect of outlying values.

**Table 5**
Relative bias (%) of point and variance estimators, within parentheses, for different second stage sampling fractions $k_i/K_i$. Population with outlying values. $N = 3000$. $n = 150$.

| $k_i/K_i$ | | EL$^\ddagger$ | RML | WGEE | CL |
|---|---|---|---|---|---|
| 0.10 | $\beta_{0N}$ | 0.20 | 0.20 (-1.66) | 0.20 (-6.14) | 0.26 (-2.13) |
| | $\beta_{1N}$ | 0.17 | 0.17 (24.77) | 0.17 (-6.21) | 0.37 (-1.85) |
| | $\beta_{2N}$ | $-1.17$ | -1.16 (33.18) | -1.17 (-7.01) | -1.62 (-4.14) |
| 0.25 | $\beta_{0N}$ | 0.14 | 0.14 (2.47) | 0.14 (-3.32) | 0.27 (0.85) |
| | $\beta_{1N}$ | 0.08 | 0.08 (26.59) | 0.08 (-5.37) | 0.32 (-1.90) |
| | $\beta_{2N}$ | $-0.72$ | -0.72 (37.44) | -0.72 (-5.08) | -1.62 (-2.36) |
| 0.40 | $\beta_{0N}$ | 0.11 | 0.11 (4.61) | 0.11 (-1.86) | 0.27 (1.25) |
| | $\beta_{1N}$ | $-0.05$ | -0.05 (30.45) | -0.05 (-3.01) | 0.21 (-0.09) |
| | $\beta_{2N}$ | $-0.48$ | -0.47 (42.33) | -0.48 (-2.59) | -1.56 (-2.06) |

‡ Variances are not provided, because they are not needed for EL inference.

The relative biases (%) are given in Table 5. Since units are selected with equal probabilities, the empirical likelihood, RML and WGEE estimators are all equal because they are based upon the same estimating equation (4), with equal inclusion probabilities. The parameter $\beta_{2N}$ is slightly underestimated. The RML approach gives positively biased variances. The variances of the point estimators are mostly underestimated with the WGEE and CL approaches, especially when the sample sizes within PSUs are small.

Table 6 gives the observed coverages of confidence intervals. The EL approach has slightly better coverages overall, even when the point estimators are not normally distributed. The coverages of the WGEE and the CL confidence intervals

**Table 6**
Observed coverages (%) of 95% confidence intervals, for different second stage sampling fractions $k_i/K_i$. Population with outlying values. D'Agostino's K-squared p-values within parentheses. $N = 3000$. $n = 150$.

| $k_i/K_i$ | | EL | RML | WGEE | CL |
|---|---|---|---|---|---|
| 0.10 | $\beta_{0N}$ | 94.8 (0.01) | 94.8 (0.01) | 93.8$^\dagger$ (0.01) | 94.5$^\dagger$ (0.02) |
| | $\beta_{1N}$ | 94.4$^\dagger$ (0.00) | 97.2$^\dagger$ (0.00) | 93.9$^\dagger$ (0.00) | 94.4$^\dagger$ (0.00) |
| | $\beta_{2N}$ | 94.2$^\dagger$ (0.01) | 97.6$^\dagger$ (0.01) | 93.8$^\dagger$ (0.01) | 94.0$^\dagger$ (0.05) |
| 0.25 | $\beta_{0N}$ | 95.0 (0.01) | 95.2 (0.01) | 94.3$^\dagger$ (0.01) | 95.1 (0.05) |
| | $\beta_{1N}$ | 94.3$^\dagger$ (0.52) | 97.1$^\dagger$ (0.53) | 94.0$^\dagger$ (0.52) | 94.2$^\dagger$ (0.14) |
| | $\beta_{2N}$ | 94.5$^\dagger$ (0.04) | 97.8$^\dagger$ (0.04) | 93.7$^\dagger$ (0.04) | 94.3$^\dagger$ (0.09) |
| 0.40 | $\beta_{0N}$ | 95.0 (0.00) | 95.5$^\dagger$ (0.00) | 94.5$^\dagger$ (0.00) | 95.0 (0.00) |
| | $\beta_{1N}$ | 94.8 (0.72) | 97.2$^\dagger$ (0.74) | 94.3$^\dagger$ (0.72) | 94.7 (0.20) |
| | $\beta_{2N}$ | 95.2 (0.87) | 98.1$^\dagger$ (0.87) | 94.6 (0.87) | 94.4$^\dagger$ (0.73) |

$^\dagger$ Coverages significantly different from 95%. p-value $\leq 0.05$.

are mostly significantly different from the nominal level 95%. The biases of the variances observed in Table 5 explain the under and over coverages.

## 7.4. Inference for model parameters

In this §, a model-design based approach is considered, as in §6. The model considered is (34), as in § 7, with $\sigma_u^2 = 8$ and $\sigma_e^2 = 4$. Different distributions for $u_i$ and $e_{ij}$ are considered. The number of PSUs is $N = 2000$, with $e_{ij} \sim N(0, \text{sd} = \sigma_e)$ and $u_i \sim N(0, \text{sd} = \sigma_u)$. The cluster sizes are $K_i = \lfloor \exp(\tau_i) \rfloor$, with $\tau_i \sim N(\gamma_0 + \gamma_1 \times (B_0 + u_i), \text{sd} = 0.224)$, $\gamma_0 = 5.8$ and $\gamma_1 = -0.09$. In this case, Pfeffermann et al.'s (2006) and Kim et al.'s (2017) model assumptions hold, because $u_i$ and $e_{ij}$ are normally distributed. Consider 500 finite populations generated from (34). For each population, $n = 100$ PSUs are selected with randomized systematic sampling with $\pi_i \propto K_i$. Simple random samples of $k_i = 12$ SSUs were selected from each PSU sampled.

The empirical likelihood estimator (EL) described in §6 is compared with the standard restricted maximum likelihood (RML) estimator, Kim et al. (2017)' estimator based on an EM algorithm (EM), Kim et al. (2017)' scaled-EM estimator and Pfeffermann et al.'s (2006) MCMC estimator. The scaled-EM estimator is obtained by multiplying $\pi_{j|i}^{-1}$ by the scaling factors (7) in the maximisation step (see Kim et al., 2017, p.484). Consider 100 bootstrap samples to estimate the variances of the EM estimators (Rao, Wu and Yue, 1992). The MCMC algorithm is implemented with the package R2WinBUGS (Sturtz, Ligges and Gelman, 2005). The length of the chain is 5000, with 3000 "*burn-in*". The following flat prior were used. $\boldsymbol{B} \sim N(0, 10^6 \boldsymbol{I}_3)$, $(\gamma_0, \gamma_1)^\top \sim N(0, 10^3 \boldsymbol{I}_2)$, $\sigma_e \sim Un(0, 10^3)$, $\sigma_u \sim Un(0, 10^3)$, $\sigma_\tau \sim Un(0, 10)$ (e.g. Pfeffermann et al., 2006), where $\boldsymbol{I}_d$ denotes the $d \times d$ identity matrix.

**Table 7**
Relative biases (%) of the point estimators of infinite population parameters. $N = 2000$. $n = 100$.

| | EL | RML | EM | scaled-EM | MCMC |
|---|---|---|---|---|---|
| $B_0$ | −0.11 | −3.68 | −4.56 | −1.07 | 0.01 |
| $B_1$ | 0.01 | −0.03 | −0.03 | 0.13 | 0.01 |
| $B_2$ | 0.08 | 0.08 | 0.08 | 0.07 | 0.11 |
| $\sigma_e^2$ | 0.13$^\ddagger$ | 0.30 | −1.62 | −2.35 | 0.59 |
| $\sigma_u^2$ | −1.58$^\ddagger$ | −0.57 | 0.28 | 1.00 | 2.54 |

‡ Relative biases (%) of the method-of-moments estimators (see Appendix A).

The observed RB are given in Table 7. The intercepts are negatively biased with RML, EM and scaled-EM. EM is clearly biased. On the other hand, the scaled-EM estimator has a smaller bias. The EL and the MCMC approaches are less biased. The observed coverages of the confidence intervals and the p-values of the normality test are given in Table 8. The MCMC confidence intervals are defined by the 2.5% and 97.5% quantiles of the posterior distribution. The EM confidence intervals are the usual confidence intervals based on variance estimates. The coverages of the EL confidence intervals are not significantly different from 95%. The alternative approaches give coverages significantly

**Table 8**
Observed coverages (%) of 95% confidence intervals for infinite population parameters. D'Agostino's K-squared p-values within parentheses. $N = 2000$. $n = 100$.

| | EL | RML | EM | scaled-EM | MCMC |
|---|---|---|---|---|---|
| $B_0$ | 95.4 (0.14) | 31.8† (0.20) | 16.2† (0.22) | 88.4† (0.40) | 96.8† (0.22) |
| $B_1$ | 95.2 (0.23) | 96.0 (0.49) | 96.0 (0.45) | 95.6 (0.30) | 95.6 (0.54) |
| $B_2$ | 93.6 (0.93) | 95.2 (0.69) | 94.8 (0.74) | 94.4 (0.49) | 96.2 (0.60) |

† Coverages significantly different from 95%. p-value $\leq 0.05$.

different from 95%, for the intercept term. The negative bias of the RML and EM point estimators explains the low coverages. This is not dues to a lack of normality, because normality is not rejected in all cases.

In order to investigate the robustness against incorrect model assumption, skewed random effects are generated; that is, $u_i = \sigma_u(v_i - 0.774) \times 1.5795$, where $v_i$ are generated from a skewed normal distribution with a location= 0, scale= 1 and shape= 4. In this case, the data deviate from Pfeffermann et al.'s (2006) and Kim et al.'s (2017) model assumptions. Consider $N = 2000$ PSUs. The cluster sizes are $K_i = \lfloor \exp(\tau_i) \rfloor$, with $\tau_i \sim N(6.5 - 0.09 \times (B_0 + u_i), \text{sd} = 0.224)$.

**Table 9**
Relative biases (%) of the point estimators of infinite population parameters. Random effects following a skewed normal distribution. $N = 2000$. $n = 100$.

| | EL | RML | EM | scaled-EM | MCMC |
|---|---|---|---|---|---|
| $B_0$ | 0.00 | −3.27 | −3.61 | −0.60 | −0.20 |
| $B_1$ | 0.01 | −0.01 | 0.00 | −0.22 | 0.00 |
| $B_2$ | −0.10 | −0.07 | −0.08 | 0.00 | −0.02 |
| $\sigma_e^2$ | 0.04‡ | 0.14 | −0.96 | −1.25 | 0.43 |
| $\sigma_u^2$ | −1.78‡ | −17.27 | −0.28 | −2.66 | −14.68 |

‡ Relative biases (%) of the method-of-moments estimators (see Appendix A).

**Table 10**
Observed coverages (%) of 95% confidence intervals for infinite population parameters. Random effects with skewed normal distribution. D'Agostino's K-squared p-values within parentheses. $N = 2000$. $n = 100$.

| | EL | RML | EM | scaled-EM | MCMC |
|---|---|---|---|---|---|
| $B_0$ | 95.2 (0.22) | 36.6† (0.32) | 31.4† (0.90) | 90.6† (0.03) | 92.8 (0.34) |
| $B_1$ | 94.2 (0.17) | 97.0† (0.03) | 95.4 (0.03) | 94.4 (0.34) | 96.8† (0.02) |
| $B_2$ | 95.6 (0.39) | 96.2 (0.09) | 95.2 (0.11) | 96.2 (0.43) | 96.8† (0.18) |

† Coverages significantly different from 95%. p-value $\leq 0.05$.

The RB and observed coverages are presented in Tables 9 and 10. The empirical likelihood approach have smaller RB and coverages not significantly different from 95%. Compared to Table 7, it seems that the skewness of $u_i$ increased significantly the RB of the RML, scaled-EM and MCMC estimators of $\sigma_u^2$. On the other hand, the method-of-moments and the EM estimators of $\sigma_u^2$ are more robust. Significantly low coverages for the confidence intervals of $B_0$ for RML, EM and scaled-EM are observed, as in Table 8. For the MCMC approach, a better coverage is observed for the $B_0$, but with over-coverages for $B_1$ and $B_2$.

## 8. Educational survey data (PISA)

In this §, the EL approach proposed is applied to the 2006 PISA survey data (OECD, 2006, 2007) for the United Kingdom. The data consist of students selected from a stratified two-stage sampling design. Information on the skills and knowledge of 15-year-old students were collected. The schools are the PSUs and the students are the SSUs (OECD, 2006). §8.1 presents the result of a simulation study. The proposed approach is applied to the original data in §8.2.

## 8.1. Simulation study

These PISA survey data were inflated to create an artificial population of size $\sum_{i \in U} K_i = 396\,768$ with $N = 4016$. The cluster sizes $K_i$ range from 12 to 220. The model (34) is used, where the response variable $y_{ij}$ is the '*mathematics achievement score on average*'. The explanatory variables are '*female*' (1 if female, and 0 otherwise) and the '*socio-economic status of parents*' given by the Ganzeboom, Graaf and Treiman's (1992) index.

The number of two-stage cluster samples elected is 10 000. The first stage is a randomized systematic samples of $n = 200$ PSUs with unequal probabilities proportional to the reciprocal of the survey school weights provided in the PISA data. The reciprocal of the weights were re-scaled to avoid very small and very large selection probabilities. For the second stage, $k_i = \alpha K_i$ SSUs are selected simple random sampling, within the $i$th sample PSU, with $\alpha = 0.25$, 0.5 and 0.75. The finite population parameter is $\boldsymbol{\beta}_N = (\beta_{0N}, \beta_{1N}, \beta_{2N})^\top = (446.82, -16.81, 1.17)^\top$.

**Table 11**
Observed coverages (%) of 95% confidence intervals for different second stage sampling fractions $k_i/K_i$. 2006 PISA survey data for the United Kingdom. D'Agostino's K-squared p-values within parentheses. $N = 4016$. $n = 200$.

| $k_i/K_i$ | $\boldsymbol{\beta}_N$ | EL | RML |
|---|---|---|---|
| 0.25 | $\beta_{0N}$ | 91.0$^\dagger$ (0.08) | 87.4$^\dagger$ (0.40) |
| | $\beta_{1N}$ | 93.5$^\dagger$ (0.55) | 82.6$^\dagger$ (0.63) |
| | $\beta_{2N}$ | 88.1$^\dagger$ (0.38) | 80.7$^\dagger$ (0.93) |
| 0.50 | $\beta_{0N}$ | 93.8$^\dagger$ (0.24) | 86.0$^\dagger$ (0.36) |
| | $\beta_{1N}$ | 94.1$^\dagger$ (0.26) | 74.6$^\dagger$ (0.86) |
| | $\beta_{2N}$ | 93.0$^\dagger$ (0.72) | 74.0$^\dagger$ (0.81) |
| 0.75 | $\beta_{0N}$ | 94.6 (0.02) | 84.3$^\dagger$ (0.07) |
| | $\beta_{1N}$ | 94.5$^\dagger$ (0.51) | 68.7$^\dagger$ (0.67) |
| | $\beta_{2N}$ | 94.2$^\dagger$ (0.09) | 65.2$^\dagger$ (0.53) |

$^\dagger$ Coverages significantly different from 95%. P-value $\leq 0.05$

The observed coverages (%) of the EL and the RML confidence intervals are provided in Table 11. The RML approach has poor coverages, because the variance estimator is negatively biased. The coverages of the EL confidence intervals are closer to the nominal level than those of the RML confidence intervals.

## 8.2. Application to the original PISA survey data

In this §, the original 2006 PISA survey data for the United Kingdom are used to compare the EL, RML, EM, scaled-EM and MCMC point estimates and their p-values. The original data consists of $\sum_{i \in S} k_i = 13\,152$ students and $n = 502$ schools. The sample cluster (school) sizes $k_i$ range from 3 to 55. The reciprocal of school and student survey weights were, respectively, used as proxies for $\pi_i$ and $\pi_{j|i}$.

Consider the regression model (34), where '*mathematics achievement score on average*' is the response variable. The covariates are

*Mother-tertiary*: 1 if mother has a tertiary education, 0 otherwise
*Parent-tertiary*: 1 if parents have tertiary education, 0 otherwise
*Public*: 1 for public school, 0 otherwise
*Large-class*: 1 for class size over 25, 0 otherwise
*City*: 1 for city located schools, 0 otherwise
*Native*: 1 for first and second generations not immigrant, 0 otherwise
*Sub-nation*: 1 for Scotland, 0 otherwise

Table 12 gives the point estimates and their p-values. Wald's test statistics have been used for RML, EM, scaled-EM and MCMC. Note all the approaches give different set of point estimates. '*Public*' is significant with all the approaches except with EM. '*City*' is significant with the RML, EM, scaled-EM and MCMC. '*City*' is not significant with EL. '*Public*' is the only significant effect with EL.

The difference between EL and RML is due to the fact that RML ignores the survey weights. Point and variance estimates may be biased, because some students can be miss-represented. This may also apply to other parametric

**Table 12**
Comparison of the methods in terms of point estimates and significance p-values. Original 2006 PISA survey data for the United Kingdom. P-values within parentheses. $n = 502$.

| | EL | RML | EM | scaled-EM | MCMC |
|---|---|---|---|---|---|
| *Mother-tert.* | -1.73 (0.99) | -0.52 (0.70) | -3.24 (0.07) | -2.05 (0.40) | -0.56 (0.68) |
| *Parent-tert.* | -2.30 (0.88) | 0.26 (0.85) | -2.37 (0.22) | -2.38 (0.41) | 0.28 (0.84) |
| *Public* | -33.20 (0.00$^\dagger$) | -14.33 (0.01$^\dagger$) | -6.75 (0.45) | -24.97 (0.02$^\dagger$) | -15.40 (0.00$^\dagger$) |
| *Large-class* | -18.32 (0.62) | -1.28 (0.75) | -1.35 (0.83) | -10.05 (0.09) | -2.73 (0.50) |
| *City* | -11.13 (0.57) | -8.86 (0.04$^\dagger$) | -21.98 (0.00$^\dagger$) | -18.05 (0.01$^\dagger$) | -8.52 (0.04$^\dagger$) |
| *Native* | 5.46 (0.25) | 7.67 (0.02$^\dagger$) | 7.63 (0.09) | 5.56 (0.31) | 7.43 (0.02$^\dagger$) |
| *Sub-nation* | 7.89 (0.96) | 9.82 (0.04$^\dagger$) | 16.95 (0.00$^\dagger$) | 6.74 (0.19) | 9.62 (0.04$^\dagger$) |

$^\dagger$ p-value $\leq 0.05$.

approaches as observed in the simulation study (see §7.4). The RML and MCMC provide similar estimates. The MCMC approach assumes that the selection probabilities of schools are proportional to schools sizes, approximated by a log-normal distribution (see §7.4 for more details). However, the selection model may not be correct. The EL point estimator takes the design into account without the need of a selection model. It also provides accurate p-values, because they are based on an ancillary test statistic which accounts for the design.

## 9. Discussion

The simulation studies shows that the empirical approach may provide less biased point estimates and better confidence intervals, even when the point estimator is not normal, the data are skewed or include outlying values. Alternative confidence intervals may have coverages significantly different from the nominal value, when sample sizes are not large enough or data include outlying values. Parametric approaches may not necessarily be robust when the model assumptions do not hold.

The empirical likelihood confidence intervals are data driven and have the advantage of not requiring variance estimates, re-sampling, linearisation and second order inclusion probabilities. They are not based on the normality of the point estimator. The empirical likelihood approach proposed takes into account the sampling design. It is also the less computer-intensive approach. The approach proposed can be extended to multilevel models with random slopes and/or complex correlation structures.

The empirical likelihood approach proposed accommodates large PSU-level sampling fraction, without the need of adjustment factors based on eigenvalues (Berger, 2018a; Zhao and Wu, 2018; Rao and Scott, 1981), design effects (Wu and Rao, 2006) or finite population corrections (Berger and Torres, 2016). With small or large sampling fraction, the pivotal property holds, when the model parameter is the target parameter. When the finite population parameter is the target, it is necessary to assume negligible sampling fraction for the pivotal property to hold, because it ensures that the variation between the model and finite population correction is negligible. Classical approaches based on variance estimates require separate estimation of a design-based variance and model-based variance. The empirical log-likelihood ratio function implicitly includes a consistent model-design-based variance estimate with an adjustment for the model variance, without the need of adjustments involving joint-inclusion probabilities (see §6). Therefore empirical likelihood confidence intervals can be constructed for model parameters, taking into account of the model-design variability, even with large sampling fractions. For example, this cannot be achieved easily with rescaled bootstrap, because it requires small sampling fractions. Traditional parametric approaches often relies on small sampling fractions.

Full response is assumed. Imputation and weight trimming are other features that exist with survey data. These problems are beyond the scope. However, the approach proposed can be combined with with Berger's (2018a) adjusted empirical likelihood approach for unit nonresponse under cluster sampling.

## Acknowledgements

Melike Oğuz-Alper was supported by the Economic and Social Research Council, United Kingdom. The authors wish to thank the reviewers and Professor Li-Chun Zhang (University of Southampton) for helpful comments. They are also grateful to an anonymous reviewer who suggested adding §6.

## Supplementary material

In Appendix A describes the estimators of the variance components. A comparative simulation study related to these estimators is also presented. The regularity conditions and the detailed proof of Theorem 1 and Lemma 1 can be found in Appendix B.

## References

Aires, N., 2000. Comparisons between conditional poisson sampling and pareto $\pi$ps sampling designs. Journal of Statistical Planning and Inference 82, 1–15.

Asparouhov, T., 2006. General multi-level modelling with sampling weights. Communication in Statistics - Theory and Methods 35, 439–460.

Berger, Y.G., 1998. Rate of convergence to asymptotic variance for the Horvitz-Thompson estimator. Journal of Statistical Planning and Inference 74, 149–168.

Berger, Y.G., 2018a. An empirical likelihood approach under cluster sampling with missing observations. Annals of the Institute of Statistical Mathematics doi:10.1007/s10463-018-0681-x.

Berger, Y.G., 2018b. Empirical likelihood approaches in survey sampling. The Survey Statistician 78, 22–31.

Berger, Y.G., 2018c. Empirical likelihood approaches under complex sampling designs. Wiley StatsRef: Statistics Reference Online , 20pp.

Berger, Y.G., 2018d. Online supplementary materials of Berger (2018a). Annals of the Institute of Statistical Mathematics , 13pp.

Berger, Y.G., Torres, O.D.L.R., 2012. A unified theory of empirical likelihood ratio confidence intervals for survey data with unequal probabilities. Proceedings of the Survey Research Method Section of the American Statistical Association, Joint Statistical Meeting, San Diego , 15URL: http://www.asasrms.org/Proceedings(Nov.2019).

Berger, Y.G., Torres, O.D.L.R., 2014. Empirical likelihood confidence intervals: an application to the EU-SILC household surveys. Contribution to Sampling Statistics, Contribution to Statistics: F. Mecatti, P. L. Conti, M. G. Ranalli (editors). Springer , 65–84.

Berger, Y.G., Torres, O.D.L.R., 2016. An empirical likelihood approach for inference under complex sampling design. Journal of the Royal Statistical Society Series B, doi: 10.1111/rssb.12115 78, 319–341.

Binder, D.A., 1983. On the variance of asymptotically normal estimators from complex surveys. Int. Stat. Rev. 51, 279–292.

Binder, D.A., Patak, Z., 1994. Use of estimating functions for estimation from complex surveys. Journal of the American Statistical Association 89, 1035–1043.

Chaudhuri, S., Handcock, M.S., Rendall, M.S., 2008. Generalized linear models incorporating population level information: An empirical-likelihood-based approach. Journal of the Royal Statistical Society Series B 70, 311–328.

Chen, J., Sitter, R.R., 1999. A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. Statist. Sinica 9, 385–406.

Chen, J., Sitter, R.R., Wu, C., 2002. Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys. Biometrika 89, 230–237.

Chen, S., Kim, J.K., 2014. Population empirical likelihood for nonparametric inference in survey sampling. Statist. Sinica 24, 335–355.

Clogg, C., Eliason, S., 1987. Some common problems in log-linear analysis. Sociological Methods and Research 16, 8–44.

Deville, J.C., 1999. Variance estimation for complex statistics and estimators: linearization and residual techniques. Survey Methodology 25, 193–203.

Deville, J.C., Särndal, C.E., 1992. Calibration estimators in survey sampling. Journal of the American Statistical Association 87, 376–382.

Diggle, P., Heagerty, P., Liang, K., Zeger, S., 2002. Analysis of longitudinal data (2nd ed.). Oxford University Press, Oxford.

D'Agostino, R., 1970. Transformation to normality of the null distribution of $g_1$. Biometrika 57, 679–681.

Fuller, W.A., 2009. Some design properties of a rejective sampling procedure. Biometrika 96, 933–944.

Ganzeboom, H.B., Graaf, P.M.D., Treiman, D.J., 1992. A standard international socio-economic index of occupational status. Social Science Research 21, 1 – 56.

Godambe, V.P., Thompson, M., 2009. Estimating functions and survey sampling, in: Pfeffermann, D., Rao, C. (Eds.), Sample Surveys: Inference and Analysis. Elsevier, Amsterdam. Handbook of Statistics, pp. 83–101.

Goldstein, H., 1986. Multilevel mixed linear model analysis using iterative generalised least squares. Biometrika 73, 43–56.

Graubard, B., Korn, E., 1996. Modelling the sampling design in the analysis of health surveys. Statistical Methods in Medical Research 5, 263–281.

Grilli, L., Pratesi, M., 2004. Weighted estimation in multilevel ordinal and binary models in the presence of informative sampling designs. Survey Methodology 30, 93–103.

Hájek, J., 1964. Asymptotic theory of rejective sampling with varying probabilities from a finite population. The Annals of Mathematical Statistics 35, 1491–1523.

Hájek, J., 1981. Sampling from a Finite Population. Marcel Dekker, New York.

Hansen, M.H., Hurwitz, W.N., 1943. On the theory of sampling from finite populations. The Annals of Mathematical Statistics 14, pp. 333–362.

Hartley, H.O., Rao, J.N.K., 1962. Sampling with unequal probabilities without replacement. The Annals of Mathematical Statistics 33, 350–374.

Haziza, D., Mecatti, F., Rao, J.N.K., 2004. Comparison of variance estimators under Rao-Sampford method: a simulation study, in: Proc. Survey Methods Sec., Am. Statist. Assoc.. pp. 3638–3643.

Horvitz, D.G., Thompson, D.J., 1952. A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association 47, 663–685.

Isaki, C.T., Fuller, W.A., 1982. Survey design under the regression super-population model. Journal of the American Statistical Association 77, 89–96.

Kim, J.K., 2009. Calibration estimation using empirical likelihood in survey sampling. Statistica Sinica 19, 145–157.

Kim, J.K., Park, S., Lee, Y., 2017. Statistical inference using generalized linear mixed models under informative cluster sampling. Canadian Journal of Statistics 45, 479–497.

Kovačević, M., Rai, S., 2003. A pseudo maximum likelihood approach to multilevel modelling of survey data. Communication in Statistics - Theory and Methods 32, 103–121.

Liang, K., Zeger, S., 1986. Longitudinal data analysis using generalized linear models. Biometrika 73, 13–22.

Longford, N., 1995. Models for uncertainty in educational testing. Springer, New York.

Matei, A., Tillé, Y., 2005. Evaluation of variance approximations and estimators in maximum entropy sampling with unequal probability and fixed sample size. Journal of Official Statistics 21, 543–570.

Neyman, J., 1938. On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. Journal of the Royal Statistical Society 97, 558–625.

OECD, 2006. Pisa 2006 technical report.

OECD, 2007. PISA 2006: Science Competencies for Tomorrow's World, Volume 1 - Analysis. OECD Publisher, Paris.

Oğuz-Alper, M., Berger, Y.G., 2016. Empirical likelihood approach for modelling survey data. Biometrika 103, 447–459.

Owen, A.B., 1988. Empirical likelihood ratio confidence intervals for a single functional. Biometrika, doi: 10.1093/biomet/75.2.237 75, 237–249.

Owen, A.B., 1991. Empirical likelihood for linear models. Ann. Statist. 19, 1725–1747.

Owen, A.B., 2001. Empirical Likelihood. Chapman & Hall, New York.

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

Pfeffermann, D., Skinner, C., Holmes, D., Goldstein, H., Rasbash, J., 1998. Weighting for unequal selection probabilities in multilevel models. Journal of the Royal Statistical Society. Series B 60, 23–40.

Potthoff, R., Woodbury, M., Manton, K., 1992. "Equivalent sample size" and "equivalent degrees of freedom" refinements for inference using survey weights under superpopulation models. Journal of the American Statistical Association 87, 383 – 396.

R Core Team, 2019. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL: `https://www.r-project.org/`.

Rabe-Hesketh, S., Skrondal, A., 2006. Multilevel modelling of complex survey data. Journal of Royal Statistical Society: Series A 169, 805–827.

Rao, J., Scott, A., 1981. The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness of fit and independence in two-way tables. Journal of the American Statistical Association 76, 221–230.

Rao, J.N.K., Molina, I., 2015. Small Area Estimation. 2nd ed., Wiley, Hoboken, NJ.

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

Rao, J.N.K., Wu, C.F.J., Yue, K., 1992. Some recent work on resampling methods for complex surveys. Survey Methodology 18, 209–217.

Särndal, C.E., Swensson, B., Wretman, J.H., 1992. Model Assisted Survey Sampling. Springer-Verlag, New York.

Skinner, C., 1989. Domain means, regression and multivariate analysis. In Analysis of Complex Surveys. C.J. Skinner, D. Holt and T.M.F. Smith (editors). Chichester: Wiley. , 59–87.

Skinner, C.J., Vieira, M.D.T., 2007. Variance estimation in the analysis of clustered longitudinal survey data. Survey Methodology 33, 3–12.

Sturtz, S., Ligges, U., Gelman, A., 2005. R2winbugs: A package for running winbugs from r. Journal of Statistical Software 12, 1–16. URL: `http://www.jstatsoft.org`.

Tan, Z., Wu, C., 2015. Generalized pseudo empirical likelihood inferences for complex surveys. The Canadian Journal of Statistics 43, 1âĂŞ17.

Wu, C., Rao, J.N.K., 2006. Pseudo-empirical likelihood ratio confidence intervals for complex surveys. Canadian Journal of Statistics, doi: 10.1002/cjs.5550340301 34, 359–375.

You, Y., Rao, J., 2002. A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights. Canadian Journal of Statistics 30, 431–439.

Zhao, P., Wu, C., 2018. Some theoretical and practical aspects of empirical likelihood methods for complex surveys. International Statistical Review 87, 239–256.