Empirical likelihood inference for regression parameters when modelling hierarchical complex survey data

Melike Oguz-Alper* Yves G. Berger†

Abstract
The data used in social, behavioural, health or biological sciences may have a hierarchical structure due to the natural structure in the population of interest or due to the sampling design. Multilevel or marginal models are often used to analyse such hierarchical data. The data may include sample units selected with unequal probabilities from a clustered and stratified population. Inferences for the regression coefficients may be invalid when the sampling design is informative. We apply a profile empirical likelihood approach to the regression parameters, which are defined as the solutions of a generalised estimating equation. The effect of the sampling design is taken into account. This approach can be used for point estimation, hypothesis testing and confidence intervals for the sub-vector of parameters. It asymptotically provides valid inference for the finite population parameters under a set of regularity conditions. We consider a two–stage sampling design, where the first stage units may be selected with unequal probabilities. We assume that the model and sampling hierarchies are the same. We treat the first stage sampling units as the unit of interest, by using an ultimate cluster approach. The estimating functions are defined at the ultimate cluster level of the hierarchy.

Key Words: Design–based inference, generalised estimating equation, empirical likelihood, two–stage sampling, uniform correlation structure, regression coefficient, unequal inclusion probability

1. Introduction

The data may be collected from samples that are selected from a multi–stage sampling design that may involve unequal probabilities at some or all stages of the selection. The sampling design is called informative when the selection probabilities are associated with the model outcome variable, even after conditioning on the model covariates. Ignoring informative sampling may result in invalid inference for regression parameters (e.g. Pfeffermann et al., 1998). Several methods that take sampling weights into account have been considered for hierarchical data (e.g. Pfeffermann et al., 1998; Asparouhov, 2006; Rabe-Hesketh and Skrondal, 2006; Skinner and De Toledo Vieira, 2007; Rao et al., 2013).

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

We propose using the profile empirical likelihood approach proposed by Oguz-Alper and Berger (2016) to make inferences for hierarchical regression parameters. We consider a sample weighted *generalised estimating equation* (GEE) (e.g. Skinner and De Toledo Vieira, 2007) to estimate regression parameters. We assume that the model and the design have the same hierarchical structure. We use an ultimate cluster approach (Hansen et al., 1953). The

*Statistics Norway, Postboks 8131 Dep, 0033, Oslo, Norway, Melike.Oguz.Alper@ssb.no. This research is funded by the Economic and Social Research Council, United Kingdom

†University of Southampton, Southampton Statistical Sciences Research Institute, SO17 1BJ, Southampton, United Kingdom, Y.G.Berger@soton.ac.uk

empirical likelihood function is defined at the ultimate cluster level by assuming that the sampling fraction is negligible at that level. The resulting empirical likelihood confidence intervals may be better than the standard confidence intervals even when the point estimator is not normal, the variance estimators are biased or unstable. The confidence intervals proposed do not rely on re–sampling, linearisation, variance estimation or design effect. Population level information can be accommodated with the approach proposed.

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 the population U is stratified into a finite number H of strata denoted by $U_1, \dots, U_H$. We have $\cup_{h=1}^{H} U_h = U$ and $\sum_{h=1}^{H} N_h = N$, where N_h denotes the number of PSUs within U_h.

Let s_h be the sample of U_i, selected *with replacement* with unequal probabilities p_i (Hansen and Hurwitz, 1943) from U_h, where $\sum_{i \in U_h} p_i = 1$. Let n_h denotes the fixed number of draws from U_h. We assume that the sampling fractions n_h/N_h are negligible. The overall sample of PSUs is $s = \cup_{h=1}^{H} s_h$. Let

$$\pi_i := n_h \, p_i \cdot$$

The samples s_h can be also a without-replacement set of units, because sampling with and without replacement are asymptotically equivalent when n_h/N_h are negligible (Hájek, 1981, p.112). Under sampling without replacement, π_i are the inclusion probabilities.

Let s_i be the sample of *secondary sampling units* (SSUs), of size k_i, with $j = 1, \dots, k_i$, selected with conditional probabilities $\pi_{j|i}$ within the ith PSU selected at the first stage. Let $\boldsymbol{v}_{ij}$ be the vector of variables associated with unit $j \in U_i$.

Suppose that we know a population parameter $\boldsymbol{\varphi}_N$ (Chaudhuri et al., 2008) which is the solution of the estimating equation

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

For example, this could the estimating equation of means or ratios. The vector $\boldsymbol{\varphi}_N$ will be treated as a vector of constant, 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 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). We assume that $n \to \infty$, where n is the number of PSUs sampled. We assume that the size of the PSUs are bounded; that is, $\max\{K_i : \ i \in U\} = O(1)$. The number of strata H, n_h/n and N_h/N are fixed constants that they do not vary as $n \to \infty$. The sampling fraction n/N is assumed negligible: $n/N = o(1)$.

3. Multilevel model of interest

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

$$y_{ij} = \boldsymbol{x}_{ij}^\top \boldsymbol{B} + \epsilon_{ij}, \tag{1}$$

where

$$\epsilon_{ij} := u_i + e_{ij}\cdot$$

Here, u_i and the e_{ij} are independent random variables with means zero and variances σ_u^2 and σ_e^2 respectively. The response variables y_{ij} are conditionally independent given the random effect u_i and marginally correlated within cluster i. This implies that the variance of $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{iK_i})^\top$, with respect to (1), 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 $K_i \times 1$ column vector of ones (e.g. Rao, 2003, p.135).

The finite population parameter $\boldsymbol{\beta}_N$ is defined as the *generalised least square* predictor of $\boldsymbol{B}$. The parameter $\boldsymbol{\beta}_N$ is the solution to the population GEE (e.g. Liang and Zeger, 1986) given by

$$\boldsymbol{G}(\boldsymbol{\beta}) := \sum_{i\in U} \boldsymbol{g}_i(\boldsymbol{\beta}) = \mathbf{0}_b, \tag{2}$$

where

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

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 the model parameter $\boldsymbol{B}$. The resulting estimator is fully efficient when the working model is correctly specified. The covariance structure $\boldsymbol{\Sigma}_i$ within $\boldsymbol{g}_i(\boldsymbol{\beta})$ does not affect the consistency but only the efficiency (Liang and Zeger, 1986; Diggle et al., 2002).

We treat $\boldsymbol{\beta}_N$ as the parameter of interest. Hence we consider a design–based inference for $\boldsymbol{\beta}_N$, where the sampling distribution is only specified by the sampling design. Under this framework, $\boldsymbol{v}_{ij}$ are treated as fixed, non-random constant vectors.

4. Sample weighted GEE estimator

The sample weighted GEE estimator $\widehat{\boldsymbol{\beta}}$ is defined as the solution to

$$\sum_{i\in s} \pi_i^{-1}\, \widehat{\boldsymbol{g}}_i(\boldsymbol{\beta}) = \mathbf{0}_b, \tag{3}$$

where

$$\widehat{\boldsymbol{g}}_i(\boldsymbol{\beta}) := \widehat{\boldsymbol{X}}_i^\top\, \widehat{\boldsymbol{\Sigma}}_i^{-1}\big(\widehat{\boldsymbol{y}}_i - \widehat{\boldsymbol{X}}_i\boldsymbol{\beta}\big)\cdot \tag{4}$$

Here, $\widehat{\boldsymbol{y}}_i$ and $\widehat{\boldsymbol{X}}_i$ are the sample-based sub-matrices of $\boldsymbol{y}_i$ and $\boldsymbol{X}_i$, which contains the observations of the sample s_i. We propose using the following estimator of $\boldsymbol{\Sigma}_i^{-1}$

$$\widehat{\boldsymbol{\Sigma}}_i^{-1} \quad := \quad \widehat{\sigma}_e^{-2}\left\{\mathrm{diag}(w_{j|i} : j \in s_i) - \widehat{\gamma}_i\, \widehat{w}_{\cdot|i}^{-1}\,(\widehat{\boldsymbol{w}}_i\, \widehat{\boldsymbol{w}}_i^\top)\right\},$$

where

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

where a_i are scaling factors. Scaling may reduce the bias of the estimators when cluster sample sizes are small (e.g. Pfeffermann et al., 1998). We consider the scaling method, so–called *scaling method 1* by Pfeffermann et al. (1998, p.30). Here, $\widehat{\sigma}_e^2$ and $\widehat{\sigma}_u^2$ are sample based estimates of σ_e^2 and σ_u^2. In the numerical work in Section 6, we consider the *method–of–moments* type of estimators to estimate σ_e^2 and σ_u^2 (e.g. Prasad and Rao, 1990; Graubard and Korn, 1996; Huang and Hidiroglou, 2003; Korn and Graubard, 2003). Poor estimation of the variance components may result in some loss in efficiency of the inference for the finite population parameter $\boldsymbol{\beta}_N$. However, the consistency will still hold provided the number of sample PSUs, n, is large (e.g. Pfeffermann et al., 1998).

The design-consistency of $\widehat{\boldsymbol{\beta}}$ can be established by using a Taylor expansion of (3) and assuming that

$$\sum_{i \in s} \pi_i^{-1}\, \widehat{\boldsymbol{g}}_i(\boldsymbol{\beta}_N) = O_p(n^{-\frac{1}{2}}),$$

(e.g. Godambe and Thompson, 2009, p.90).

5. Empirical likelihood approach

Consider the *empirical log-likelihood function* (Berger and De La Riva Torres, 2016) given by

$$\ell(\boldsymbol{m}) = \sum_{i \in s} \log m_i, \tag{5}$$

where the m_i are unknown scale-loads allocated to data points $i \in s$ (Hartley and Rao, 1968) and $\boldsymbol{m}$ denotes the $n \times 1$ vector of m_i. Here, the m_i are defined for the PSUs sampled.

Let $\widehat{m}_i^*(\boldsymbol{\beta})$ maximizes $\ell(\boldsymbol{m})$ subject to the constraints $m_i > 0$ and

$$\sum_{i \in s} m_i\, \boldsymbol{c}_i^*(\boldsymbol{\beta}) = \boldsymbol{C}^*, \tag{6}$$

with

$$\begin{aligned} \boldsymbol{c}_i^*(\boldsymbol{\beta}) &:= \{\boldsymbol{c}_i^\top, \widehat{\boldsymbol{g}}_i(\boldsymbol{\beta})^\top\}^\top, & \boldsymbol{C}^* &= (\boldsymbol{C}^\top, \boldsymbol{0}^\top)^\top, \\ \boldsymbol{c}_i &:= \{\bar{\pi}^{-1}\boldsymbol{z}_i^\top, \widehat{\mathbf{f}}_i^\top\}^\top, & \boldsymbol{C} &= (\bar{\pi}^{-1}\boldsymbol{n}^\top, \boldsymbol{0}^\top)^\top, \\ \boldsymbol{z}_i &:= (z_{i1}, \ldots, z_{iH})^\top, & \boldsymbol{n} &:= \sum_{i \in U} \boldsymbol{z}_i = (n_1, \ldots, n_H)^\top, \\ \widehat{\mathbf{f}}_i &:= \sum_{j \in s_i} \pi_{j|i}^{-1} \mathbf{f}_{ij}. \end{aligned}$$

where $z_{ih} = \pi_i$ for $i \in U_h$ and $z_{ih} = 0$ otherwise. Here, $\bar{\pi} = n/N$ is the sampling fraction at PSU level, $\widehat{\boldsymbol{g}}_i(\boldsymbol{\beta})$ is defined by (4) and $\mathbf{f}_{ij} := \mathbf{f}(\boldsymbol{\varphi}_N, \boldsymbol{v}_{ij})$. The solution to this maximisation is invariant to $\bar{\pi}$, because $\bar{\pi}$ appears on both side of the constraint (6). Hence, the approach can be implemented with $\bar{\pi}$ removed from both $\boldsymbol{c}_i$ and $\boldsymbol{C}$. The quantity $\bar{\pi}$ is only required to ensure that $\boldsymbol{c}_i$ is bounded, which is a necessary requirement to justify the asymptotic results.

We assume that the $\boldsymbol{C}^*$ is an inner point of the conical hull formed by $\sum_{i\in s} m_i \boldsymbol{c}_i^*(\boldsymbol{\beta})$, with $\boldsymbol{\beta}$ in the parameter space. Hence the set of $\widehat{m}_i^*(\boldsymbol{\beta})$ is unique.

The maximum value of $\ell(\boldsymbol{m})$ under $m_i > 0$ and (6) is given by

$$\ell(\boldsymbol{\beta}) = \sum_{i\in s} \log \widehat{m}_i^*(\boldsymbol{\beta})\cdot \tag{7}$$

This function takes into account of the sampling design and the population level information because of $\boldsymbol{z}_i$ and $\widehat{\mathbf{f}}_i$ within the definition of $\boldsymbol{c}_i^*(\boldsymbol{\beta})$.

The *maximum empirical likelihood estimator* $\widehat{\boldsymbol{\beta}}_{EL}$ of $\boldsymbol{\beta}_N$ is the vector that maximizes expression (7). It can be shown that $\widehat{\boldsymbol{\beta}}_{EL}$ is the solution of the following sample level estimating equation (Berger and De La Riva Torres, 2016)

$$\widehat{\boldsymbol{G}}(\boldsymbol{\beta}) = \sum_{i\in s} \widehat{m}_i\, \widehat{\boldsymbol{g}}_i(\boldsymbol{\beta}) = \mathbf{0}_b, \tag{8}$$

where the $\widehat{m}_i$ are the *maximum empirical likelihood weights* obtained by maximising the empirical log-likelihood function (5) with respect to the constraints: $m_i > 0$ and

$$\sum_{i\in s} m_i \boldsymbol{c}_i = \boldsymbol{C}; \tag{9}$$

that is,

$$\widehat{m}_i = (\pi_i + \boldsymbol{\eta}^\top \boldsymbol{c}_i)^{-1}, \tag{10}$$

where the vector $\boldsymbol{\eta}$ is such that (9) and $m_i > 0$ hold. A modified Newton–Raphson algorithm as in Chen et al. (2002) can be used to compute $\boldsymbol{\eta}$.

When we do not use any population level information, we have $\boldsymbol{c}_i = \bar{\pi}^{-1}\boldsymbol{z}_i$. In this case, $\boldsymbol{\eta} = \mathbf{0}$ and $\widehat{m}_i = \pi_i^{-1}$, the standard Horvitz and Thompson's (1952) weight for the ith PSU. In this case, $\widehat{\boldsymbol{\beta}}_{EL}$ is the weighted GEE estimator $\widehat{\boldsymbol{\beta}}$, which is the solution to (3).

Suppose that the parameter of interest $\boldsymbol{\theta}_N$ is a sub-parameter of $\boldsymbol{\beta}_N$; that is, $\boldsymbol{\beta}_N = (\boldsymbol{\theta}_N^\top, \boldsymbol{\nu}_N^\top)^\top$, where $\boldsymbol{\nu}_N$ is a nuisance parameter. We wish to make inference about $\boldsymbol{\theta}_N$ in the presence of the nuisance parameters $\boldsymbol{\nu}_N$. The profile empirical log-likelihood ratio function is defined by

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

where $\ell(\boldsymbol{\theta},\boldsymbol{\nu}) = \ell(\boldsymbol{\beta})$ with $\boldsymbol{\beta} = (\boldsymbol{\theta}^\top, \boldsymbol{\nu}^\top)^\top$ and $\ell(\widehat{\boldsymbol{\beta}}) = \sum_{i\in s} \log \widehat{m}_i$, where the $\widehat{m}_i$ are defined by (10). Under some regularity conditions (Oguz-Alper and Berger, 2016) and using an ultimate cluster approach (Hansen et al., 1953), we have that

$$\widehat{r}(\boldsymbol{\theta}_N) \xrightarrow{d} \chi^2_{df=p} \tag{11}$$

in distribution, with respect to the sampling design. Thus $\widehat{r}(\boldsymbol{\theta}_N)$ is pivotal and can be used for testing hypotheses. Confidence intervals can be constructed based on (11), when $\boldsymbol{\theta}_N$ is scalar.

6. Simulation study

In this Section, we present some numerical results for the parameters of a hierarchical linear model defined by (1). Our simulation study shows the design performance of confidence intervals. We selected 10 000 random samples with respect to a two-stage sampling design from a finite population, which is a realisation of the model

$$y_{ij} = B_0 + B_1 x_{ij}^{(1)} + B_2 x_{ij}^{(2)} + u_i + e_{ij}, \tag{12}$$

where $B_0 = 20$, $B_1 = B_2 = 1$, $x_{ij}^{(1)}$ and $x_{ij}^{(2)}$ are generated according to gamma distributions; that is, $x_{ij}^{(1)} \sim \mathrm{gamma}(K_i, \mathrm{shape} = 2, \mathrm{scale} = \alpha_{1i})$ and $x_{ij}^{(2)} \sim \mathrm{gamma}(K_i, \mathrm{shape} = 2, \mathrm{scale} = \alpha_{2i})$, where α_{1i} and α_{2i} are selected randomly with replacement among the values $(1, 2, 3)$ and $(1, 2, 3, 4)$ respectively. The K_i are the cluster sizes generated from a log-normal distribution and defined by $K_i = 100 \exp(\tau_i)$, with $\tau_i \sim N(0, 0.2)$. The number of clusters is $N = 3000$. The cluster sizes range between 47 and 207. The random effects u_i follow a normal distribution with mean zero and standard deviation σ_u. The e_{ij} are the level one residuals generated from a chi-squared distribution, that is, $e_{ij} \sim \chi^2(\sigma_e^2/2) - \sigma_e^2/2$, with $\sigma_e^2 = 12 - \sigma_u^2$. The values of σ_u^2 were chosen in such a way that different intra–cluster correlations, defined by $\rho = \sigma_u^2/\sigma^2$, where $\sigma^2 = \sigma_e^2 + \sigma_u^2 = 12$, were obtained. The total variance, σ^2, was kept fixed at 12. The correlation coefficients considered range from 0.04 to 0.83. Population size is $\sum_{i \in U} K_i = 305\,305$. The finite population parameter, $\boldsymbol{\beta}_N = (\beta_{0N}, \beta_{1N}, \beta_{2N})^\top$, is obtained by solving (2) and given in Table 1.

We selected two-stage samples. At the first stage, a sample s of $n = 150$ PSUs was selected with randomized systematic sampling with unequal probabilities π_i proportional to $\delta_i = b_0 + u_i + b_1 \epsilon_i$, where $\epsilon_i \sim \exp(rate = 1) - 1$ and $\mathrm{corr}(\delta_i, u_i) \approx 0.85$. The constant b_0 was used to avoid very small inclusion probabilities. The constant b_1 was used to control the correlation between δ_i and u_i. For the second stage, samples of $k_i = \alpha K_i$ SSUs were selected with simple random sampling without replacement within the PSU i selected, where $\alpha = 0.25$. The range of sample sizes within clusters is $[12, 52]$.

We compare the Monte Carlo design–based performance of the empirical likelihood confidence interval with the confidence intervals obtained from the restricted maximum likelihood, the weighted GEE (e.g. Skinner and De Toledo Vieira, 2007) and the composite likelihood (Rao et al., 2013) approaches. The nominal level considered is 95%. The weighted GEE and the composite likelihood confidence intervals rely on variance estimates. We used the Hartley and Rao (1962) variance estimator for the former and the conditional variance estimator (e.g. Rao et al., 2013, p.270) for the latter.

The weighted GEE confidence interval is based on the method of inverse testing (e.g. Binder, 1983; Binder and Patak, 1994), which takes the randomness of the nuisance parameters into account. The composite likelihood confidence interval relies on the linearised sandwich variance estimator (e.g. Rao et al., 2013, p.270). Normality is implicitly assumed.

Standard parametric confidence interval involves restricted maximum likelihood estimation. The point estimator is assumed to be normal. Hierarchical structure is considered by fitting a two–level model with a uniform covariance structure. Survey weights are not taken into account with this approach. Point estimates and standard errors were obtained by using the 'lme' function in R (R Development Core Team, 2014).

We used the D'Agostino's (1970) K–squared test for the test of normality of the point estimators. We also considered the percentage relative bias (%) of the point estimators with respect to the sampling design. The percentage relative bias (%) is defined by $\mathrm{RB}\% = [\{E(\widehat{\phi}) - \phi\}/\phi] * 100\%$, where $E(\widehat{\phi}) = M^{-1} \sum_{m=1}^{M} \widehat{\phi}_m$, with $M = 10\,000$, is the empirical expectation of the estimator $\widehat{\phi}$, where ϕ is the parameter of interest and $\widehat{\phi}_m$ is an estimate based on the mth sample.

We consider three estimators denoted by $\widehat{\boldsymbol{\beta}}^{rml}$, $\widehat{\boldsymbol{\beta}}^{cl}$ and $\widehat{\boldsymbol{\beta}}^{el}$ in Table 1, where $\widehat{\boldsymbol{\beta}}^{rml} = (\widehat{\beta}_0^{rml}, \widehat{\beta}_1^{rml}, \widehat{\beta}_2^{rml})^\top$ is the restricted maximum likelihood estimator, $\widehat{\boldsymbol{\beta}}^{cl} = (\widehat{\beta}_0^{cl}, \widehat{\beta}_1^{cl}, \widehat{\beta}_2^{cl})^\top$ is the composite likelihood estimator and $\widehat{\boldsymbol{\beta}}^{el} = (\widehat{\beta}_0^{el}, \widehat{\beta}_1^{el}, \widehat{\beta}_2^{el})^\top$ is the empirical likelihood estimator proposed. Weights are incorporated with the empirical, weighted GEE and composite likelihood approaches. The estimator $\widehat{\boldsymbol{\beta}}^{el}$ is the solution to the sample GEE (8). We obtain the same point estimator with the empirical likelihood and the weighted GEE approaches, because there is no population level information, and thus the weights are equivalent to the Horvitz and Thompson (1952) weights. The restricted maximum likelihood estimator $\widehat{\boldsymbol{\beta}}^{rml}$ does not involve weights.

In Table 1, we have the finite population values of regression coefficients and the relative bias (%) of their estimators. The restricted maximum likelihood point estimator $\widehat{\beta}_0^{rml}$ is slightly biased, because the selection of the PSUs is informative due to the dependency on random effects given in (12). Relative biases (%) increase with intra–cluster correlation.

Table 1: Finite population values of the regression coefficients in working model (12) and the relative bias (%) of their estimators. Two–stage sampling design. Unequal probability selection of the PSUs. $N = 3000$ and $n = 150$.

Intra–cluster correlation	Population value			Relative bias (%) *unweighted* (RML)			*weighted* (CL)			*weighted* (EL)		
ρ	β_{0N}	β_{1N}	β_{2N}	$\widehat{\beta}_0^{rml}$	$\widehat{\beta}_1^{rml}$	$\widehat{\beta}_2^{rml}$	$\widehat{\beta}_0^{cl}$	$\widehat{\beta}_1^{cl}$	$\widehat{\beta}_2^{cl}$	$\widehat{\beta}_0^{el}$	$\widehat{\beta}_1^{el}$	$\widehat{\beta}_2^{el}$
0.04	20.00	1.00	1.00	0.63	0.01	-0.07	0.05	0.07	-0.10	0.03	0.00	-0.06
0.25	20.05	1.00	1.00	2.05	0.03	0.00	0.03	0.50	-0.13	0.02	0.03	0.03
0.50	20.06	1.00	1.00	2.62	0.00	-0.01	0.07	0.55	-0.19	0.01	0.03	-0.03
0.83	20.06	1.00	1.00	3.22	0.02	-0.03	0.10	0.74	-0.26	0.03	0.01	-0.01

RML, CL and EL stand respectively for restricted maximum likelihood, composite likelihood and empirical likelihood.

In Table 2, we have the observed coverages of the 95% confidence intervals. The D'Agostino's K–squared test of normality shows that the point estimators are mostly not normally distributed. Coverages of the restricted maximum likelihood confidence intervals for the intercept are significantly different from the nominal level, 95%, for all cases. The coverages with this approach decreases with intra–cluster correlation. The poor coverages are due to ignoring the informative sampling. In this case, the finite population parameter β_{0N} is slightly over–estimated and the variance of $\widehat{\beta}_0^{rml}$ is substantially underestimated. We observe poor coverages with the weighted GEE and composite likelihood approaches in most cases. The empirical likelihood confidence intervals have better coverages overall.

7. Conclusion

We proposed using an empirical likelihood approach to make design–based inference for regression parameters when modelling hierarchical data. The approach proposed provides asymptotically valid design–based inference. The numerical work show that the empirical likelihood confidence intervals may provide better coverages than the standard confidence intervals based on the normality assumption, even when the point estimator is not normal or the data is skewed. Standard confidence intervals that do not take sampling design into account may have very poor coverages. The empirical likelihood approach proposed does not depend on variance estimation, re–sampling, linearisation and second order inclusion probabilities. Confidence intervals are not based on the normality of the point estimator.

Table 2: 95% confidence intervals for the estimates of regression coefficients. Two–stage sampling design. Unequal probability selection of the PSUs. $N = 3000$ and $n = 150$. The D'Agostino's K–squared p-values within the parentheses.

Intra–cluster correlation ρ	Parameter $\boldsymbol{\beta}_N$	Empirical likelihood (%)	Restricted ML (%)	Weighted GEE (%)	Composite likelihood (%)
0.04	β_0	94.9 (0.89)	82.9*(0.60)	94.3*(0.89)	94.6 (0.90)
	β_1	94.8 (0.00)	95.0 (0.00)	94.2*(0.00)	94.8 (0.00)
	β_2	94.9 (0.71)	94.9 (0.56)	94.2*(0.71)	94.5*(0.93)
0.25	β_0	94.6 (0.00)	31.4*(0.33)	94.1*(0.00)	94.2*(0.02)
	β_1	94.8 (0.79)	94.9 (0.56)	94.3*(0.79)	93.8*(0.00)
	β_2	94.1*(0.01)	94.7 (0.01)	93.7*(0.01)	94.2*(0.03)
0.50	β_0	95.1 (0.00)	28.7*(0.13)	94.6 (0.00)	94.7 (0.01)
	β_1	94.6 (0.05)	95.1 (0.18)	94.0*(0.05)	93.9*(0.84)
	β_2	94.6 (0.00)	95.4 (0.00)	94.1*(0.00)	94.2*(0.00)
0.83	β_0	94.9 (0.11)	27.7*(0.63)	94.3*(0.11)	94.5*(0.09)
	β_1	94.7 (0.03)	95.4 (0.11)	94.1*(0.03)	94.3*(0.63)
	β_2	94.5*(0.01)	94.8 (0.00)	93.9*(0.01)	94.5*(0.05)

$*$ Coverages significantly different from 95%. p-value ≤ 0.05.

REFERENCES

Asparouhov, T. (2006), "General multi-level modelling with sampling weights," *Communication in Statistics - Theory and Methods*, 35(3), 439–460.

Berger, Y. G., and De La Riva Torres, O. (2016), "An empirical likelihood approach for inference under complex sampling design," *Journal of the Royal Statistical Society Series B*, 78(2), 319–341.

Binder, D. A. (1983), "On the variance of asymptotically normal estimators from complex surveys," *Int. Stat. Rev.*, 51(427), 279–292.

Binder, D. A., and Patak, Z. (1994), "Use of estimating functions for estimation from complex surveys," *Journal of the American Statistical Association*, 89(427), 1035–1043.

Chaudhuri, S., Handcock, M. S., and 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(2), 311–328.

Chen, J., Sitter, R. R., and Wu, C. (2002), "Using Empirical Likelihood Methods to Obtain Range Restricted Weights in Regression Estimators for Surveys," *Biometrika*, 89(1), 230–237.

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

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

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

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

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.

Hájek, J. (1981), *Sampling from a Finite Population*, New York: Marcel Dekker.

Hansen, M. H., and Hurwitz, W. N. (1943), "On the Theory of Sampling from Finite Populations," *The Annals of Mathematical Statistics*, 14(4), pp. 333–362.

Hansen, M., Hurwitz, W., and Madow, W. (1953), *Sample Survey Methods and Theory, volume I*, New York: John Wiley and Sons.

Hartley, H. O., and Rao, J. N. K. (1962), "Sampling with unequal probabilities without replacement," *Ann. math. Statist. Assoc.*, 33, 350–374.

Hartley, H. O., and Rao, J. N. K. (1968), "A new estimation theory for sample surveys," *Biometrika*, 55(3), 547–557.

Horvitz, D. G., and Thompson, D. J. (1952), "A Generalization of Sampling Without Replacement From a Finite Universe," *Journal of the American Statistical Association*, 47(260), 663–685.

Huang, R., and Hidiroglou, M. (2003), “Design consistent estimators for a mixed linear model on survey data,” *Proceedings of the Survey Research Method Section of the American Statistical Association, Joint Statistical Meetings, San Francisco*, pp. 1897–1904.

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

Korn, E., and Graubard, B. (2003), “Estimating variance components by uisng survey data,” *Journal of the Royal Statistical Society. Series B*, 65, 175–190.

Liang, K., and Zeger, S. (1986), “Longitudinal data analysis using generalized linear models,” *Biometrika*, 73, 13–22.

Oguz-Alper, M., and Berger, Y. G. (2016), “Empirical likelihood approach for modelling survey data,” *Biometrika*, 103(2), 447–459.

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

Prasad, N., and Rao, J. (1990), “The estimation of the mean square error of small-area estimators,” *Journal of American Statistical Association*, 85, 163–171.

R Development Core Team (2014), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing. http://www.R-project.org, Vienna, Austria.

Rabe-Hesketh, S., and Skrondal, A. (2006), “Multilevel modelling of complex survey data,” *Journal of Royal Statistical Society: Series A*, 169(4), 805–827.

Rao, J. N. K. (2003), *Small Area Estimation*, : Wiley, Hoboken, NJ.

Rao, J. N. K., Verret, F., and Hidiroglou, M. (2013), “A weighted composite likelihood approach to inference for two-level models from survey data,” *Survey Methodology*, 39(2), 263–282.

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.*, pp. 59–87.

Skinner, C. J., and De Toledo Vieira, M. (2007), “Variance estimation in the analysis of clustered longitudinal survey data,” *Survey Methodology*, 33(1), 3–12.