Unconditional empirical likelihood approach for analytic use of public survey data

YVES G. BERGER

Southampton Statistical Sciences Research Institute, University of Southampton

Abstract

Modelling survey data often requires having the knowledge of design and weighting variables. With public-use survey data, some of these variables may not be available for confidentiality reasons. The proposed approach can be used in this situation, as long as calibrated weights and variables specifying the strata and primary sampling units are available. It gives consistent point estimation and a pivotal statistics for testing and confidence intervals. The proposed approach does not rely on with-replacement sampling, single-stage, negligible sampling fractions or non-informative sampling. Adjustments based on design effects, eigenvalues, joint-inclusion probabilities or bootstrap, are not needed. The inclusion probabilities and auxiliary variables do not have to be known. Multi-stage designs with unequal selection of primary sampling units are considered. Non-response can be easily accommodated if the calibrated weights include re-weighting adjustment for non-response. We use an unconditional approach, where the variables and sample are random variables. The design can be informative.

Key Words: calibration, estimating equation, informative sampling, multi-stage sampling, unequal inclusion probability, weights

Running Headline: Empirical likelihood for public surveys

1. Introduction

Fitting models on complex public-use survey data for secondary data analysis is common practice in many areas of socials sciences and economics. Survey data are rarely composed of independent and identically distributed (i.i.d.) observations, because the data collection process, called sampling design, often involves clustering, stratification and unequal inclusion probability selection (e.g. Skinner *et al.*, 1989; Chambers & Skinner, 2003). Estimation often relies on adjustments for the sampling design. However, users of public-use survey data often have limited design information, given by calibrated weights and stratification/clustering variables. With secondary data analysis, inclusion probabilities and the auxiliary variables used for weighting are often not available because of confidentiality.

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/sjos.12590

Suppose that the data are selected with multi-stage unequal probabilities designs, with small or large sampling fractions. The proposed unconditional approach has the advantage of taking the sampling design into account and the model defining the parameter. It does not need the auxiliary variables and inclusion probabilities. The point estimator is consistent. The empirical log-likelihood ratio function follows a χ^2-distribution asymptotically, under the null hypothesis, without adjustment involving eigenvalues, design effects, variance estimates, finite population corrections or bootstrap.

We assume that the public-use survey data contain some calibration weights derived from auxiliary variables and inclusion probabilities (e.g. Deville & Särndal, 1992). Theses weights may also include some non-response adjustments. There are numerous situations the inclusion probabilities and auxiliary variables are unavailable to data users, because they contain sensitive information, which may identify some units. For example, the *European Union Statistics on Income and Living Conditions* (Eurostat, 2012) users' databases do not contain auxiliary variables. Another example, is when the values of non-surveyed auxiliary variables are linked to sampled units, for weighting purpose. These auxiliary variables cannot be released by the producer of the data, when interviewees have not given their consent to make them publicly available.

The weights allows consistent point estimation. However, design-based variance estimates implicitly rely on (first and joint) inclusion probabilities and auxiliary variables (Deville & Särndal, 1992) or some bootstrap weights. Consistent design-based variance estimation may not be possible if this information is not available. The aim of this paper is to show that under the proposed approach, this information is not required.

When the sampling design is non-informative (or ignorable), we can use a parametric approach based a conditional "*sample-likelihood*", given the sample labels (e.g. Chambers *et al.*, 2012). We may also use a disaggregated model with random effects to control for clustering and dichotomous variables for stratification. However, the inclusion probabilities can be associated with some variables of interest, i.e. the design could be informative or non-ignorable. In this case, the sample-likelihood needs to be adjusted for the sampling design, using the Bayes's theorem (e.g. Krieger & Pfeffermann, 1992; Pfeffermann *et al.*, 1998; Pfeffermann & Sverchkov, 1999). In other words, the information about the design is incorporated within a likelihood-based framework, by specifying a model for the relationship between the inclusion probabilities and some variables of interest. This implicitly assumes that these probabilities are known, which is not the case for public-use data files. The adjusted sample-likelihood account for informativeness, but relies on re-

strictive assumptions about the design, such as "*asymptotic independence*" which is achieved under sampling with-replacement, Poisson sampling or negligible sampling fractions. The distribution of a sample-likelihood ratio test statistics cannot be easily derived, and variance estimation and tests are often based on bootstrap.

Informative sampling is challenging with a conditional "*sample-likelihood*" framework. We shall see that with the proposed unconditional approach, informative sampling is naturally accounted for within the empirical likelihood function, without requiring a model for the inclusion probabilities, distributional assumptions or limitation on the design. Testing can be based on the empirical log-likelihood ratio function. Furthermore, the sample likelihood requires making assumptions about the unknown distribution of error terms, and unknown heteroscedasticity, which is not an issue with the proposed approach. Furthermore, the framework considered has the advantage of not depending on the correlation structure of error terms, and can therefore accommodate intra-PSU correlations.

The proposed approach shares some common features with pseudo-likelihood (Binder & Roberts, 2009), based on sample-based weighted estimating equations which estimate unbiasedly a population score function. This should not be confused with the mainstream pseudo-likelihood approach (Besag, 1975), and should be viewed as a weighted version of the "*generalized method of moments*". For testing, Wald or F statistics based on variance estimates can be used (see Appendix C in the supplement). Rao & Scott (1981) proposed using a naïve Wald test statistics adjusted by eigenvalues computed from variance estimates, based inclusion probabilities and auxiliary variables, which are assumed unknown in our setup. Empirical likelihood tests are usually more powerful than Wald-tests.

Empirical likelihood for survey data comes in different flavours called: "*pseudoempirical likelihood*" (Chen & Sitter, 1999; Wu & Rao, 2006), "*unequal probability empirical likelihood*" (Berger & Torres, 2012, 2014, 2016) and "*population empirical likelihood*" (Chen & Kim, 2014). An approach based on biased Poisson sampling was consider by Kim (2009). They all based on a conditional design-based (non-parametric) framework. The "*sample empirical likelihood*" of Chen & Kim (2014) is a particular case of the approach of Berger & Torres (2012; 2014; 2016). A Bayesian version based on unequal probability empirical likelihood, can be found in Zhao *et al.* (2019). Wu & Rao (2006) proposed adjusting the pseudoempirical log-likelihood ratio function with design effects, when the parameter is scalar. Zhao & Wu (2018) and Berger (2018) proposed adjusting the unequal probability empirical log-likelihood ratio function with eigenvalues to account for the design, similar in spirit to Rao & Scott (1981) and Wu & Rao (2006). Variance estimation based on inclusion probabilities is needed for the adjustment pro-

posed by Chen & Sitter (1999), Wu & Rao (2006), Zhao & Wu (2018) and Berger (2018). They are also conditional (design-based) approaches. The empirical likelihood proposed approach is unconditional and motivated by the unequal probability empirical likelihood. The empirical log-likelihood ratio function does not need to be adjusted with variance estimates, inclusion probabilities, weighting auxiliary variables, eigenvalues, design effect or bootstrap. It is solely based on calibration weights and variables specifying the stratification and clustering. This proposed approach is asymptotically valid even when the within cluster sample sizes are small.

Chaudhuri & Handcock (2018) proposed a "*conditional empirical likelihood*" technique; which can be viewed as an empirical likelihood version of the fully parametric sample-likelihood procedure of Pfeffermann *et al.* (1998). It relies on assumptions about the design such conditional independence and a model for the inclusion probabilities, as in Pfeffermann *et al.* (1998). This implicitly means that the inclusion probabilities have to be known. The distribution of the conditional empirical likelihood ratio statistics is not available. Wald test statistics is proposed. Our approach is different and wider in scope. It is unconditional and is not motivated by assumptions about the design. Our main contribution is to provide the asymptotic distribution of the empirical log-likelihood ratio statistics.

If a matrix of re-scaled bootstrap weights (Rao *et al.*, 1992) is made available with a public-use survey dataset, it can be used for point and variance estimation. Zhao *et al.* (2020) exploited this idea to derive an empirical likelihood approach under public-use survey data, which contain bootstrap weights. Zhao *et al.* (2020) proposed adjusting the pseudoempirical or empirical likelihood ratio function with eigenvalues computed from a bootstrap variance estimate as in Wu & Rao (2006), Berger (2018) and Zhao & Wu (2018). They also suggest a "*bootstrap calibration method*" which consists in computing the bootstrap quantiles of the distribution of the empirical log-likelihood ratio statistics under with-replacement sampling. This method can be extremely computer intensive, since computing a single values of a empirical log-likelihood ratio function is also intensive, because it involves optimisation. For the bootstrap calibration method, the bootstrap weights need to be produced in a very specific way and may be different from the re-scaled bootstrap weights available. The information about the design and auxiliary information is in-bedded within the matrix of bootstrap weights. This is a elegant and sensible technique for inference. However, there are several issues. Organisations releasing public-use data do not always provide this matrix of bootstrap weights, which cannot be constructed by users, if the inclusion probabilities and auxiliary variable are not available. Zhao *et al.* (2020) assumed that the bootstrap variance is design-consistent. However, bootstrap is valid under sampling with-replacement

or negligible sampling fractions (Rao *et al.*, 1992), which is not always the case with social data. For instance, non negligible sampling fractions are used with the "*National Health and Nutrition Examination Survey*" (National Center for Health Statistics, 2016). In the real data example of Section 9, the sampling fraction is approximately 10%. Zhao *et al.* (2020) target the finite population predictor of $\boldsymbol{\theta}_0$ under a conditional design-based approach. In this paper, we consider a unconditional model-based approach, with $\boldsymbol{\theta}_0$ being the parameter of interest. When modelling survey data, it is more natural to target the model parameter $\boldsymbol{\theta}_0$, and base the inference on the model and the design when it is informative. Zhao *et al.* (2020) bootstrap approach has the advantage of taking the effect of the auxiliary variables into account, when bootstrap weights are available and the bootstrap variance is consistent, i.e. when the sampling fraction are small or under with-replacement sampling. This affect is not taken into account with the proposed approach in this paper. Nevertheless, we allow for large sampling fractions and sampling without-replacement, without the need of design effects, eigenvalues, joint-inclusion probabilities or bootstrap weights.

In Section 2, we define the class of models considered. In Section 3 and Section 4, we define the class of sampling designs and the survey weights. The asymptotic framework and some of the regularity conditions are outlined in Section 5. The empirical likelihood approach is introduced in Section 6. The main contribution can be found in Section 7, which shows the consistency of the variance within the self-normalised profile empirical log-likelihood ratio function. In Section 8, a simulation study supports our findings. The proposed approach is illustrated using the "*Programme for International Student Assessment*" (PISA) survey data in Section 9. It shows that we may fit very different models, if we use the empirical likelihood proposed approach rather than a random effect method. Additional regularity conditions are provided in Appendix A. Computational algorithms and proofs can be find in the supplement.

2. Class of models specified by moment conditions

Let $\boldsymbol{Y} \in \mathbb{R}^{d_y}$ denote a random vector, which usually contains response variables, some covariates and some additional side variables. We consider a class of models specified by "*moment conditions*", i.e. let $\boldsymbol{\theta}_0 \in \mathbb{R}^{d_\theta}$ be a parameter specified by

$$\mathbb{E}_{\mathcal{M}}\big\{\boldsymbol{g}(\boldsymbol{Y},\boldsymbol{\theta})\big\} = \mathbf{0}_{d_g}, \qquad \text{if and only if } \boldsymbol{\theta} = \boldsymbol{\theta}_0, \tag{1}$$

where $\boldsymbol{g}(\boldsymbol{Y},\boldsymbol{\theta}) \in \mathbb{R}^{d_g}$ is an estimating function, with $d_g \geqslant d_\theta$. Here, $\mathbb{E}_{\mathcal{M}}$ is the expectation with respect to a model, and $\mathbf{0}_d$ denotes an d-vector of 0. Equation

(1) specifies a wide class of models, which includes generalised linear models and non-linear (in the parameter) models. It has the advantage of not relying on the distribution of error terms. Sample data will be used for estimating $\boldsymbol{\theta}_0$.

With endogenous covariates, the proposed approach can be used by incorporating a suitable instrument matrix with the definition of $\boldsymbol{g}(\boldsymbol{Y}, \boldsymbol{\theta})$ (e.g. Donald *et al.*, 2009), by implicitly assuming that the instrument identifies the parameter (Newey, 1993). However, Domínguez & Lobato (2004) pointed out that this may not be always the case. Berger & Patilea (2022) proposed an empirical likelihood approach that deals with this identification problem, with survey data.

We may also have some optional "*side information*". Suppose that we know a vector $\boldsymbol{d}_0$ which is the solution to another moment condition, i.e.

$$\mathbb{E}_{\mathcal{M}}\big\{\mathbf{f}(\boldsymbol{Y}, \boldsymbol{d})\big\} = \mathbf{0}_{d_{\mathbf{f}}}, \qquad \text{if and only if } \boldsymbol{d} = \boldsymbol{d}_0, \tag{2}$$

where $\mathbf{f}(\boldsymbol{Y}, \boldsymbol{d}) \in \mathbb{R}^{d_{\mathbf{f}}}$. Here, $\boldsymbol{d}_0$ is treated as known, rather than as a parameter to estimate. The vector $\boldsymbol{d}_0$ called "*side information*", usually takes the form of descriptive statistics of some of the variables within $\boldsymbol{Y}$ (e.g. Chaudhuri *et al.*, 2008). For example, $\boldsymbol{d}_0$ could be aggregate information from a census or large surveys, such as vector of means, totals, ratios or quantiles, so that it can be assumed that $\boldsymbol{d}_0$ is vector of constants. Side information can also be used in microeconometric (Imbens & Lancaster, 1994) when combining micro and macro data. Taking into account of (2) may improve the estimation of $\boldsymbol{\theta}_0$ (e.g. Deville & Särndal, 1992; Imbens & Lancaster, 1994; Chaudhuri *et al.*, 2008).

The distributions of $\boldsymbol{g}(\boldsymbol{Y}, \boldsymbol{\theta}_0)$ and $\mathbf{f}(\boldsymbol{Y}, \boldsymbol{d}_0)$ do not need to be specified. We simply assume that their second-order moments exists, i.e.

$$\mathbb{E}_{\mathcal{M}}\big\{\|\boldsymbol{g}(\boldsymbol{Y}, \boldsymbol{\theta}_0)\|^2\big\} < \infty, \tag{3}$$

$$\mathbb{E}_{\mathcal{M}}\big\{\|\mathbf{f}(\boldsymbol{Y}, \boldsymbol{d}_0)\|^2\big\} < \infty\cdot \tag{4}$$

Similar fourth-order moment conditions will be also needed, such as condition (A.4) in Appendix A.

3. Clustered population, sampling design and weighting

Suppose that we have a population of M units clustered into N primary sampling units (PSUs) denoted $\mathcal{U}_1, \ldots, \mathcal{U}_i, \ldots, \mathcal{U}_N$. Let $U := \{1, \ldots, N\}$ be the set of labels of N PSUs. Suppose that a sample S of n PSUs is selected without-replacement with unequal probabilities π_i. The randomly selected sample S is a random variable. We assume that S is a "*stratified sample*" of PSUs from H strata. The quantity

n_h denotes the number of PSUs sampled from strata h, where $h = 1, \ldots, H$ and $\sum_{h=1}^{H} n_h = n$. Within each PSU $\mathcal{U}_i$ sampled ($i \in S$), a sample S_i of n_i units is selected. The overall number of units selected is $m = \sum_{i \in S} n_i$.

We have a two-stage design when S_i is a single stage design. If the S_i contain more than two-stages, we have a multi-stage design. We have a single-stage cluster designs, when $S_i = \mathcal{U}_i$. Furthermore, if the PSUs $\mathcal{U}_i$ are made up of a single unit, we have a single-stage design. Without loss of generality, hereafter we shall consider two-stage designs. Non-response can be an ultimate phase.

We impose no restriction on the sampling fractions n_h/N_h which may be large or small. We assume that the first stage sample sizes n_h are not random and fixed by design.

We consider an unconditional framework, where both the sample and the population values $\{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_M\}$ are random variables. The first-order inclusion probabilities are also assumed random. Inference will not be based on the conditional sampling distribution given the sample's labels or $\{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_M\}$, as with model-based or design-based approaches.

Let

$$\boldsymbol{g}_{j|i}(\boldsymbol{\theta}) := \boldsymbol{g}(\boldsymbol{Y}_j, \boldsymbol{\theta}) \qquad \text{for } j \in \mathcal{U}_i.$$

Here, $\boldsymbol{Y}_j$ is the observed value of $\boldsymbol{Y}$, for unit $j \in \mathcal{U}_i$. We shall assume that the random variables $\boldsymbol{g}_{j|i}(\boldsymbol{\theta})$ are independent between PSUs, i.e.

$$\boldsymbol{g}_{j|i}(\boldsymbol{\theta}_0) \perp\!\!\!\perp \boldsymbol{g}_{\ell|k}(\boldsymbol{\theta}_0) \qquad \forall\, j \in \mathcal{U}_i,\ \ell \in \mathcal{U}_k \text{ and } i \neq k. \tag{5}$$

This assumption allows for possible correlation within PSUs, i.e. $\boldsymbol{g}_{j|i}(\boldsymbol{\theta}_0)$ and $\boldsymbol{g}_{\ell|i}(\boldsymbol{\theta}_0)$ can be dependent.

For example, suppose that the response variable follows a random effect model given by $Y_j = \boldsymbol{x}_{j|i}^{\top}\boldsymbol{\theta}_0 + \epsilon_{j|i}$, for $j \in \mathcal{U}_i$, with $\epsilon_{j|i} = u_i + e_{ji}$ with u_i and e_{ji} both independent and identically distributed. Here, $\boldsymbol{x}_{j|i}$ represents some exogenous covariates. Now the estimating function is $\boldsymbol{g}_{j|i}(\boldsymbol{\theta}_0) = \boldsymbol{x}_{j|i}(Y_j - \boldsymbol{x}_{j|i}^{\top}\boldsymbol{\theta}_0) = \boldsymbol{x}_{j|i}\,(u_i + e_{ji})$. We see that we have an intra-PSU correlation between the $\boldsymbol{g}_{j|i}(\boldsymbol{\theta}_0)$ and $\boldsymbol{g}_{\ell|i}(\boldsymbol{\theta}_0)$, because of the random effect u_i. Nonetheless, $\boldsymbol{g}_{j|i}(\boldsymbol{\theta}_0)$ and $\boldsymbol{g}_{\ell|k}(\boldsymbol{\theta}_0)$ are independent for $i \neq k$, as in (5). With the proposed approach, it will not be necessary to specify a random effect and its distribution.

4. Survey weights

Let $w_{j|i}$ denote some survey weights of a unit $j \in S_i$. We assume that there exists some quantities $\sigma_{j|i}$ such that

$$w_{j|i} = \frac{\sigma_{j|i}}{\pi_i \pi_{j|i}}, \tag{6}$$

where $\pi_{j|i}$ is the probability of selecting a unit $j \in S_i$ and π_i is the probability of selecting PSU i. The $\sigma_{j|i}$ can be g-weights (Särndal *et al.*, 1992, §6.5) or some calibration adjustment (Deville & Särndal, 1992) derived from auxiliary variables. Assumption (6) means that the weights $w_{j|i}$ have been derived by adjusting the sampling weights $(\pi_i \pi_{j|i})^{-1}$ by some re-weighting techniques. The quantities $\sigma_{j|i}$ depend on the units in the entire sample. We suppose that the $w_{j|i}$ are known for $j \in S_i$ and $i \in S$. However, $\sigma_{j|i}$, π_i and $\pi_{j|i}$ are not available to the secondary users (or data analysts). This corresponds to the usual situation when some weights $w_{j|i}$ are provided as part of public-use data files and the more sensitive information given by the inclusion probabilities and calibration variables are not revealed.

Hereafter, we shall use Greek letters for unknown quantities and Latin letters for known quantities. We treat $\sigma_{j|i}$, π_i and $\pi_{j|i}$ as random variables. The probabilities π_i and $\pi_{j|i}$ may be associated with $\boldsymbol{Y}$ when the design is informative. The function $\mathbf{f}(\boldsymbol{Y}, \boldsymbol{d})$ used within (2) could be, but does not have to be based upon the auxiliary variables used for computing the weights $w_{j|i}$..

Consider two "*weighted estimating functions*" for PSU i:

$$\breve{\boldsymbol{g}}_i(\boldsymbol{\theta}) := \sum_{j \in S_i} w_{j|i}\, \boldsymbol{g}_{j|i}(\boldsymbol{\theta}), \tag{7}$$

$$\widehat{\boldsymbol{\rho}}_i(\boldsymbol{\theta}) := \pi_i\, \breve{\boldsymbol{g}}_i(\boldsymbol{\theta}) = \sum_{j \in S_i} \frac{\sigma_{j|i}}{\pi_{j|i}} \boldsymbol{g}_{j|i}(\boldsymbol{\theta}). \tag{8}$$

The key feature is the fact that the weighted functions $\breve{\boldsymbol{g}}_i(\boldsymbol{\theta})$ are known and can be computed for a given value of $\boldsymbol{\theta}$. On the other hand, the $\widehat{\boldsymbol{\rho}}_i(\boldsymbol{\theta})$ cannot be computed by a secondary users (or data analysts), because the π_i are not available to them.

5. Asymptotic framework

The asymptotic framework is outlined as follows. We considers a sequence of nested populations of $N_{[t]}$ PSUs, with $0 < N_{[t]} < N_{[t+1]}$, and a sequence of samples of $n_{[t]}$ PSUs, with $\forall t$, $0 < n_{[t]} < n_{[t+1]}$ and $n_{[t]} < N_{[t]}$, with $t \to \infty$ (e.g. Hájek, 1964; Isaki & Fuller, 1982). To simplify notation, the index t will be dropped in what follows. Thus, $t \to \infty$ implies that $N \to \infty$, $n \to \infty$, and $m \to \infty$. We

assume that n/N, n_h/n and N_h/N and H are constants free of the limiting process, where N_h denotes the number of PSUs within stratum h. Hence, $n_h \to \infty$ and $N_h \to \infty$. We also consider that $n_h - N_h \to \infty$. This excludes heavily stratified design where n_h is bounded and H tends to infinity. The PSU sizes are considered bounded asymptotically, i.e. $m/n < \infty$ and $M/N < \infty$. The within-PSU sizes are assumed finite. Thus, the proposed approach is valid asymptotically even when the cluster sample sizes are small. Let $o_p(\cdot)$ and $O_p(\cdot)$ be the order of convergence in probability with respect to the model and the sampling design.

We assume

$$\frac{1}{N}\Big\|\sum_{i\in S}\breve{\boldsymbol{g}}_i(\boldsymbol{\theta}_0)\Big\| = O_p(n^{-\frac{1}{2}}), \tag{9}$$

$$\frac{1}{N}\Big\|\sum_{i\in S}\mathbb{E}_d^{(2)}\{\breve{\boldsymbol{g}}_i(\boldsymbol{\theta}_0)\mid S\}\Big\| = O_p(n^{-\frac{1}{2}}), \tag{10}$$

where $\|\cdot\|$ denotes the Frobenius norm. Here, $\mathbb{E}_d^{(2)}$ is the expectation with respect to the second-stage design. Analogous assumptions about the side information are given by (A.1) and (A.2) in Appendix A. Conditions (9) and (10) state that the law-of-large-numbers holds for $\breve{\boldsymbol{g}}_i(\boldsymbol{\theta}_0)$ and $\mathbb{E}_d^{(2)}\{\breve{\boldsymbol{g}}_i(\boldsymbol{\theta}_0)\mid S\}$. These are standard assumptions which are often made for deriving asymptotic properties of survey estimators (e.g. Fuller, 2009 §6).

Let $\boldsymbol{\xi}_{j|i}$ be the unknown vector of auxiliary variables used for producing $w_{j|i}$, with $j \in S_i$. The Result 5 of Deville & Särndal (1992) implies that $N^{-1}\sum_{i\in S}\pi_i^{-1}\widehat{\boldsymbol{\rho}}_i(\boldsymbol{\theta}_0)$ can be approximated by a regression estimator, i.e.

$$\begin{aligned}\frac{1}{N}\sum_{i\in S}\frac{1}{\pi_i}\widehat{\boldsymbol{\rho}}_i(\boldsymbol{\theta}_0) = \frac{1}{N}\sum_{i\in S}\sum_{j\in S_i} w_{j|i}\,\boldsymbol{g}_{j|i}(\boldsymbol{\theta}_0) = \frac{1}{N}\sum_{i\in S}\sum_{j\in S_i}\frac{1}{\pi_i\pi_{j|i}}\boldsymbol{g}_{j|i}(\boldsymbol{\theta}_0)\\ - \widehat{\boldsymbol{\beta}}_{\boldsymbol{\xi g}}^{\top}\frac{1}{N}\Big(\sum_{i\in S}\sum_{j\in S_i}\frac{1}{\pi_i\pi_{j|i}}\boldsymbol{\xi}_{j|i} - \xi_{\mathcal{U}}\Big) + o_p(n^{-\frac{1}{2}}),\end{aligned} \tag{11}$$

where $\xi_{\mathcal{U}} := N^{-1}\sum_{i\in\mathcal{U}}\sum_{j\in\mathcal{U}_i}\boldsymbol{\xi}_{j|i}$,

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{\xi g}} := \Big(\sum_{i\in S}\sum_{j\in S_i}\frac{q_{j|i}\boldsymbol{\xi}_{j|i}\boldsymbol{\xi}_{j|i}^{\top}}{\pi_i\pi_{j|i}}\Big)^{-1}\sum_{i\in S}\sum_{j\in S_i}\frac{q_{j|i}\boldsymbol{\xi}_{j|i}\boldsymbol{g}_{j|i}(\boldsymbol{\theta}_0)^{\top}}{\pi_i\pi_{j|i}} \tag{12}$$

is a regression coefficient and $q_{j|i}$ are some factors used for deriving $w_{j|i}$. Let us assume that

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{\xi g}} = o_p(1), \tag{13}$$

$$\frac{1}{N}\Big(\sum_{i\in S}\sum_{j\in S_i}\frac{1}{\pi_i\pi_{j|i}}\boldsymbol{\xi}_{j|i} - \xi_{\mathcal{U}}\Big) = O_p(n^{-\frac{1}{2}})\cdot \tag{14}$$

The key assumption is (13), which holds under normal circumstance when the auxiliary variables $\boldsymbol{\xi}_{j|i}$ are uncorrelated with $\boldsymbol{g}_{j|i}(\boldsymbol{\theta}_0)$. This is a situation usually met in practice, when $\boldsymbol{g}_{j|i}(\boldsymbol{\theta}_0)$ is a estimating function of a generalised linear model. Since $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\xi g}}$ is the regression coefficient between $\boldsymbol{\xi}_{j|i}$ and $\boldsymbol{g}_{j|i}(\boldsymbol{\theta}_0)$, and $\boldsymbol{\xi}_{j|i}$ is unknown, a user cannot compute the vector $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\xi g}}$ and would not know if $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\xi g}}$ is indeed negligible. Therefore, it will not be possible for a user to know if (13) holds. If (13) does not hold, we expect more conservative variance estimates and confidence intervals.

When making inference about $\boldsymbol{\theta}_0$, it is impossible to take the effect of the unknown variables $\boldsymbol{\xi}_{j|i}$ into account, if no proxies or bootstrap weight are available. The only solution is to conjecture that $\boldsymbol{\xi}_{j|i}$ are not correlated with $\boldsymbol{g}_{j|i}(\boldsymbol{\theta}_0)$; in other words, to assume (13). If proxies of some auxiliary variables are available, they can always be used as side variables.

Now, (11), (13) and (14) imply that

$$\frac{1}{N}\sum_{i\in S}\frac{1}{\pi_i}\widehat{\boldsymbol{\rho}}_i(\boldsymbol{\theta}_0) = \frac{1}{N}\sum_{i\in S}\frac{1}{\pi_i}\widehat{\boldsymbol{\rho}}_{\pi,i}(\boldsymbol{\theta}_0) + o_p(n^{-\frac{1}{2}}), \tag{15}$$

where

$$\widehat{\boldsymbol{\rho}}_{\pi,i}(\boldsymbol{\theta}) := \sum_{j\in S_i}\frac{1}{\pi_{j|i}}\boldsymbol{g}_{j|i}(\boldsymbol{\theta})\cdot \tag{16}$$

6. PSU-level empirical likelihood

The "*maximum empirical likelihood estimator*" is defined by

$$\widehat{\boldsymbol{\theta}} := \arg\max_{\boldsymbol{\theta}\in\Theta}\ell(\boldsymbol{\theta}), \tag{17}$$

where Θ is the parameter space and $\ell(\boldsymbol{\theta})$ is the PSU-*level empirical log-likelihood function* given by

$$\begin{aligned}\ell(\boldsymbol{\theta}) := \max_{p_i>0:\, i\in S}\Big\{\sum_{i\in S}\log(np_i) : \sum_{i\in S}p_i\breve{\boldsymbol{g}}_i(\boldsymbol{\theta}) = \mathbf{0}_{d_g},\\ n\sum_{i\in S}p_i\boldsymbol{z}_i = \overrightarrow{\boldsymbol{n}}, \sum_{i\in S}p_i\breve{\mathbf{f}}_i = \mathbf{0}_{d_{\mathbf{f}}}\Big\},\end{aligned} \tag{18}$$

where

$$\begin{aligned}\boldsymbol{z}_i &:= (z_{i1},\ldots,z_{ih},\ldots,z_{iH})^\top,\\ \overrightarrow{\boldsymbol{n}} &:= (n_1,\ldots,n_H)^\top,\\ \breve{\mathbf{f}}_i &:= \sum_{j\in S_i}w_{j|i}\,\mathbf{f}_{j|i},\\ \mathbf{f}_{j|i} &:= \mathbf{f}(\boldsymbol{Y}_j,\boldsymbol{d}_0), \quad \text{with } j\in\mathcal{U}_i\cdot\end{aligned} \tag{19}$$

Here, $z_{ih} = 1$ if PSU i belongs to strata h and $z_{ih} = 0$ otherwise. Note that the p_i within (18) are defined at PSU-level. It does not mean that we use a PSU-level model, because the model (1) is specified at unit-level. The function $\ell(\boldsymbol{\theta})$ is not a parametric likelihood, but we shall see that it behaves like a likelihood.

The Lagrange function associated with the optimisation in (18) is

$$Q(\boldsymbol{\lambda}^*, p_i \colon i \in S) := \sum_{i \in S} \log(np_i) - (\boldsymbol{t} + \boldsymbol{\lambda}^*)^\top \Big\{ n \sum_{i \in S} p_i \, \breve{\boldsymbol{c}}_i^*(\boldsymbol{\theta}) - \boldsymbol{C}^* \Big\}, \tag{20}$$

where

$$\breve{\boldsymbol{c}}_i^*(\boldsymbol{\theta}) := \Big\{ \frac{N}{n} \boldsymbol{z}_i^\top, \breve{\mathbf{f}}_i^\top, \breve{\boldsymbol{g}}_i(\boldsymbol{\theta})^\top \Big\}^\top, \qquad \boldsymbol{C}^* := \Big(\frac{N}{n} \overrightarrow{\boldsymbol{n}}^\top, \mathbf{0}_{d_{\mathbf{f}}+d_g}^\top \Big)^\top \tag{21}$$

and $\boldsymbol{t} := (\mathbf{1}_H^\top, \mathbf{0}_{d_{\mathbf{f}}+d_g}^\top)^\top$ is a scaling factor for the Lagrange multiplier $\boldsymbol{t} + \boldsymbol{\lambda}^*$. Here, $\mathbf{1}_H$ is an H-vector of 1. Since $\boldsymbol{t}^\top \breve{\boldsymbol{c}}_i^*(\boldsymbol{\theta}) = 1$, we have that maximising (20) with respect to p_i and $\boldsymbol{\lambda}^*$ reduces to a dual optimisation problem given by

$$\ell(\boldsymbol{\theta}) = -\sum_{i \in S} \log\Big\{ 1 + \widehat{\boldsymbol{\lambda}}^*(\boldsymbol{\theta})^\top \breve{\boldsymbol{c}}_i^*(\boldsymbol{\theta}) \Big\},$$

where

$$\begin{aligned} \widehat{\boldsymbol{\lambda}}^*(\boldsymbol{\theta}) &:= \arg \max_{\boldsymbol{\lambda}^* \in \Lambda^*(\boldsymbol{\theta})} \Big[\sum_{i \in S} -\log\Big\{ 1 + \boldsymbol{\lambda}^{*\top} \breve{\boldsymbol{c}}_i^*(\boldsymbol{\theta}) \Big\} - \boldsymbol{\lambda}^{*\top} \boldsymbol{C}^* \Big], \\ \Lambda^*(\boldsymbol{\theta}) &:= \Big\{ \boldsymbol{\lambda}^* \colon 1 + \boldsymbol{\lambda}^{*\top} \breve{\boldsymbol{c}}_i^*(\boldsymbol{\theta}) > n^{-1} \Big\}. \end{aligned}$$

Note that $\boldsymbol{\lambda}^* \in \Lambda^*(\boldsymbol{\theta})$ ensures that $0 < p_i \leqslant 1$, as in Qin & Lawless (1994). The computation of $\widehat{\boldsymbol{\lambda}}^*(\boldsymbol{\theta})$ is discussed in Appendix B of the supplement.

The function (18) reduces to Berger & Torres (2012, 2014, 2016) empirical likelihood function, when $\sigma_{j|i} = 1 \; \forall i, j$, where $\sigma_{j|i}$ is given by (6). Furthermore, if we have a single stratum and a single stage design, (18) becomes the sample empirical likelihood function of Chen & Kim (2014). Note that we consider $\sigma_{j|i} \neq 1 \; \forall i, j$ implicitly, because $\sigma_{j|i} = 1$ means that the π_i are known.

6.1 Empirical likelihood test

Testing will be based on profiling, as in Qin & Lawless (1994). The main contribution of the paper is to show that the profile empirical log-likelihood ratio function has a χ^2-distribution under the null (see (29)).

Let $\boldsymbol{\psi}_0 \in \mathbb{R}^{d_\psi}$ be a sub-parameter of $\boldsymbol{\theta}_0$, i.e. $\boldsymbol{\theta}_0 = (\boldsymbol{\psi}_0^\top, \boldsymbol{\upsilon}_0^\top)^\top$, where $\boldsymbol{\upsilon}_0$ is the remaining part of $\boldsymbol{\theta}_0$. Suppose that we wish to test $H_0 : \boldsymbol{\psi}_0 = \widetilde{\boldsymbol{\psi}}$ against an alternative. The test statistics we proposed, is the PSU-level "*profile empirical*

log-likelihood ratio function" defined by

$$\widehat{r}(\boldsymbol{\psi}) := 2\big\{\ell(\widehat{\boldsymbol{\theta}}) - \ell(\boldsymbol{\psi}, \widehat{\boldsymbol{\upsilon}}^{\circ})\big\}, \tag{22}$$

$\widehat{\boldsymbol{\upsilon}}^{\circ} := \arg\max_{\boldsymbol{\upsilon}\in\Upsilon} \ell(\boldsymbol{\psi}, \boldsymbol{\upsilon})$, with $\ell(\boldsymbol{\psi}, \boldsymbol{\upsilon}) := \ell(\boldsymbol{\theta})$ defined by (18), with $\boldsymbol{\theta} := (\boldsymbol{\psi}^{\top}, \boldsymbol{\upsilon}^{\top})^{\top}$. Here, $\boldsymbol{\psi}$ denotes a vector within the parameter space of $\boldsymbol{\psi}_0$ and Υ is the parameter space of $\boldsymbol{\upsilon}_0$. We shall show that $\widehat{r}(\widetilde{\boldsymbol{\psi}})$ follows an asymptotic χ^2-distribution under H_0. This result has only been shown in a more restrictive design-based framework, involving negligible sampling fractions, weighting auxiliary variables and known π_i (Oğuz-Alper & Berger, 2016a; Berger, 2018).

First, by using (8), we have that (18) reduces to

$$\ell(\boldsymbol{\theta}) = \max_{p_i>0:\, i\in S}\Big\{\sum_{i\in S}\log(np_i) : \sum_{i\in S}\frac{p_i}{\pi_i}\widehat{\boldsymbol{\rho}}_i(\boldsymbol{\theta}) = \mathbf{0}, n\sum_{i\in S}p_i\boldsymbol{z}_i = \overrightarrow{\boldsymbol{n}}, \sum_{i\in S}\frac{p_i}{\pi_i}\widehat{\boldsymbol{\phi}}_i = \mathbf{0}_{d_{\mathbf{f}}}\Big\}.$$

where $\widehat{\boldsymbol{\phi}}_i := \pi_i\breve{\mathbf{f}}_i$ is an unknown quantity. Under regularity conditions given in Appendix A, we have

$$\widehat{r}(\boldsymbol{\psi}_0) = \frac{1}{N^2}\widehat{\boldsymbol{\Gamma}}(\boldsymbol{\theta}_0)^{\top}(\mathbf{I} - \widehat{\boldsymbol{\Omega}})\widehat{\boldsymbol{\Sigma}}(\boldsymbol{\theta}_0)^{-1}\widehat{\boldsymbol{\Gamma}}(\boldsymbol{\theta}_0) + O_p(n^{-\frac{1}{2}}), \tag{23}$$

where $\widehat{\boldsymbol{\Gamma}}(\boldsymbol{\theta}_0)$ $\widehat{\boldsymbol{\Sigma}}(\boldsymbol{\theta}_0)$ and $\widehat{\boldsymbol{\Omega}}$ are defined respectively by (24), (25) and (26). The proof of (23) can be found in Oğuz-Alper & Berger (2016b) (more details can be found in Appendix D of the supplement). Here, $\widehat{\boldsymbol{\Gamma}}(\boldsymbol{\theta}_0)$ is a regression estimator defined by

$$\widehat{\boldsymbol{\Gamma}}(\boldsymbol{\theta}_0) := \sum_{i\in S}\frac{1}{\pi_i}\widehat{\boldsymbol{\rho}}_i(\boldsymbol{\theta}_0) - \widehat{\boldsymbol{B}}(\boldsymbol{\theta}_0)^{\top}\sum_{i\in S}\frac{1}{\pi_i}\widehat{\boldsymbol{\phi}}_i = \sum_{i\in S}\frac{1}{\pi_i}\widehat{\boldsymbol{\epsilon}}_i(\boldsymbol{\theta}_0), \tag{24}$$

where the residuals $\widehat{\boldsymbol{\epsilon}}_i(\boldsymbol{\theta}_0)$ and the regression coefficient $\widehat{\boldsymbol{B}}(\boldsymbol{\theta}_0)$ are given by

$$\begin{aligned}
\widehat{\boldsymbol{\epsilon}}_i(\boldsymbol{\theta}_0) &:= \widehat{\boldsymbol{\rho}}_i(\boldsymbol{\theta}_0) - \widehat{\boldsymbol{B}}(\boldsymbol{\theta}_0)^{\top}\widehat{\boldsymbol{\phi}}_i,\\
\widehat{\boldsymbol{B}}(\boldsymbol{\theta}_0) &:= \Big\{\sum_{i\in S}\pi_i^{-2}\widehat{\boldsymbol{\phi}}_i\,\widehat{\boldsymbol{\phi}}_i^{\top} - \widehat{\boldsymbol{\phi}}\,\boldsymbol{S}^{-1}\widehat{\boldsymbol{\phi}}^{\top}\Big\}^{-1}\Big\{\sum_{i\in S}\pi_i^{-2}\widehat{\boldsymbol{\phi}}_i\,\widehat{\boldsymbol{\rho}}_i(\boldsymbol{\theta}_0)^{\top} - \widehat{\boldsymbol{\phi}}\,\boldsymbol{S}^{-1}\widehat{\boldsymbol{\rho}}^{\top}\Big\},
\end{aligned}$$

where

$$\widehat{\boldsymbol{\rho}} := \sum_{i\in S}\pi_i^{-1}\widehat{\boldsymbol{\rho}}_i(\boldsymbol{\theta}_0)\,\boldsymbol{z}_i^{\top}, \quad \boldsymbol{S} := \sum_{i\in S}\boldsymbol{z}_i\boldsymbol{z}_i^{\top} \quad \text{and} \quad \widehat{\boldsymbol{\phi}} := \sum_{i\in S}\pi_i^{-1}\widehat{\boldsymbol{\phi}}_i\,\boldsymbol{z}_i^{\top}.$$

Note that we obtain the residuals $\widehat{\boldsymbol{\epsilon}}_i(\boldsymbol{\theta}_0)$ because of the side information constraint $\sum_{i\in S}p_i\breve{\mathbf{f}}_i = \mathbf{0}_{d_{\mathbf{f}}}$ within (18). The vector $\widehat{\boldsymbol{B}}(\boldsymbol{\theta}_0)$ should not be confused with (12).

The crucial term of (23) is $\widehat{\boldsymbol{\Sigma}}(\boldsymbol{\theta}_0)$, defined by

$$\widehat{\boldsymbol{\Sigma}}(\boldsymbol{\theta}_0) := \frac{1}{N^2}\sum_{i\in S}\frac{1}{\pi_i^2}\widehat{\boldsymbol{\epsilon}}_i(\boldsymbol{\theta}_0)\,\widehat{\boldsymbol{\epsilon}}_i(\boldsymbol{\theta}_0)^\top - \widehat{\boldsymbol{\tau}}, \tag{25}$$

with

$$\widehat{\boldsymbol{\tau}} := \frac{1}{N^2}\widehat{\boldsymbol{\epsilon}}\,\boldsymbol{S}^{-1}\widehat{\boldsymbol{\epsilon}}^\top \quad \text{and} \quad \widehat{\boldsymbol{\epsilon}} := \sum_{i\in S}\frac{1}{\pi_{\mathrm{i}}}\widehat{\boldsymbol{\epsilon}}_{\mathrm{i}}(\boldsymbol{\theta}_0)\boldsymbol{z}_i^\top\cdot$$

The matrix $\widehat{\boldsymbol{\Omega}}$ is defined by

$$\widehat{\boldsymbol{\Omega}} := \widehat{\boldsymbol{\Sigma}}(\boldsymbol{\theta}_0)^{-\frac{1}{2}}\widehat{\boldsymbol{\nabla}}\Big\{\widehat{\boldsymbol{\nabla}}^\top\widehat{\boldsymbol{\Sigma}}(\boldsymbol{\theta}_0)^{-1}\widehat{\boldsymbol{\nabla}}\Big\}^{-1}\widehat{\boldsymbol{\nabla}}^\top\widehat{\boldsymbol{\Sigma}}(\boldsymbol{\theta}_0)^{-\frac{1}{2}}, \tag{26}$$

where $\widehat{\boldsymbol{\nabla}} := \sum_{i\in S}\pi_i^{-1}\partial\,\widehat{\boldsymbol{\epsilon}}_i(\boldsymbol{\theta}_0)/\partial\boldsymbol{\upsilon}$. It is possible to generalise (23) to non-differentiable functions, by assuming that $\widehat{\boldsymbol{\Gamma}}(\boldsymbol{\theta}_0)$ converges to a differentiable function, as in Zhao & Wu (2018). However, this is beyond the scope of this paper, which aims to show that the empirical likelihood-based inference captures the effect of the design and the model, even with large sampling fractions.

Since $\widehat{\boldsymbol{\Gamma}}(\boldsymbol{\theta}_0)$ is a regression estimator, we have (Robinson & Särndal, 1983)

$$\frac{1}{N}\mathbb{E}_{\mathcal{M}}\mathbb{E}_d\big\{\widehat{\boldsymbol{\Gamma}}(\boldsymbol{\theta}_0)\big\} = o(n^{-\frac{1}{2}}), \tag{27}$$

where $\mathbb{E}_d$ is the expectation with respect to the sampling design. Furthermore, we assume that the central limit theorem holds for $\widehat{\boldsymbol{\Gamma}}(\boldsymbol{\theta}_0)$, i.e.

$$N^{-1}\widehat{\boldsymbol{\Gamma}}(\boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}), \tag{28}$$

where $\boldsymbol{\Sigma}$ is the "*model-design variance*" of the estimator $N^{-1}\widehat{\boldsymbol{\Gamma}}(\boldsymbol{\theta}_0)$, defined by (32) in Section 7. We shall treat (27) and (28) as regularity conditions. Justification for (28) can be found in Fuller (2009, §1.3) and Bertail *et al.* (2017).

The key result of this paper is to show that (25) is a consistent estimator of $\boldsymbol{\Sigma}$ (see Theorem 2). In this case, (23), (28) imply that under $H_0 : \boldsymbol{\psi}_0 = \widetilde{\boldsymbol{\psi}}$,

$$\widehat{r}(\widetilde{\boldsymbol{\psi}}) \xrightarrow{d} \chi^2_{df=d_\psi}, \tag{29}$$

in distribution with respect to the model and design. Here, $\chi^2_{df=d_\psi}$ denotes a χ^2-distribution with d_ψ degrees of freedom, where d_ψ is the trace d_ψ of the symmetric idempotent matrix (26). Note that equation (24) is a weighted sum of residuals $\widehat{\boldsymbol{\epsilon}}_i(\boldsymbol{\theta}_0)$, because of the constraint $\sum_{i\in S}p_i\breve{\mathbf{f}}_i = \mathbf{0}_{d_\mathbf{f}}$ within (18). Thus, the $\breve{\mathbf{f}}_i$ may reduce the variance of (24) and increase the power, because (25) is a residual vari-

ance.

Property (29) means that $\widehat{r}(\widetilde{\boldsymbol{\psi}})$ is an ancillary test statistics, which behaves like the usual likelihood ratio statistics. It can be used for testing the relative fit of two nested models, with $\boldsymbol{\psi}_0$ being the additional parameters of the full model, and $\widetilde{\boldsymbol{\psi}} = \mathbf{0}$. A goodness of fit test can also be based on (29); for example, when $\boldsymbol{\upsilon}_0$ is an intercept. Confidence intervals can be constructed when $\boldsymbol{\psi}_0$ is scalar ($d_\psi = 1$), i.e. the confidence interval with a nominal level α is

$$\mathrm{CI}(\boldsymbol{\psi}_0) := \big\{\boldsymbol{\psi} : \widehat{r}(\boldsymbol{\psi}) \leqslant \chi^2_1(\alpha)\big\}, \tag{30}$$

where $\chi^2_1(\alpha)$ is the upper α-quantile of the χ^2-distribution with 1 degree of freedom.

Lemma 1 shows that the test on $\boldsymbol{\theta}_0$ based on (22) with $\boldsymbol{\psi} = \boldsymbol{\theta}$ is consistent asymptotically against alternatives $H_a : \boldsymbol{\theta}_0 = \widetilde{\boldsymbol{\theta}}$; where $\widetilde{\boldsymbol{\theta}} := \boldsymbol{\theta}_0 + \boldsymbol{\delta} b_n$ and $\boldsymbol{\delta}$ is such that $\|\boldsymbol{\delta}\| = 1$, i.e. the local power tends to 1, as $n \to \infty$.

Lemma 1 *Under* (9)*,* (10)*, and under the conditions* (A.1)–(A.5)*,* (D.42)–(A.8) *and* (A.11) *from Appendix A, we have that there exists a sequence* $r_{\min} \to \infty$*, such that* $\mathbb{P}\{\widehat{r}(\widetilde{\boldsymbol{\theta}}) \geqslant r_{\min}\} \to 1$*, as* $n \to \infty$*, where* $\widetilde{\boldsymbol{\theta}} := \boldsymbol{\theta}_0 + \boldsymbol{\delta} b_n$*, for some* $\boldsymbol{\delta} \in \mathbb{R}^{d_\theta}$*, such that* $\|\boldsymbol{\delta}\| = 1$*. Here,* b_n *denotes an arbitrary sequence tending to zero, and such that* $nb_n^2 \to \infty$*.*

Lemma 1 implies Theorem 1 below, which establishes the $\sqrt{m}$-consistency of $\widehat{\boldsymbol{\theta}}$.

Theorem 1 *Under the conditions of Lemmas 1, we have* $m^{\frac{1}{2}}\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| = O_p(1)$.

The proof of Lemma 1 and Theorem 1 can be found in Appendix D of the supplement. Note that the asymptotic normality of $\widehat{\boldsymbol{\theta}}$ can be also established from (28).

Variance estimation is not needed for testing, since tests and confidence intervals can be based on (29). Nonetheless, an estimate for the variance matrix of $\widehat{\boldsymbol{\theta}}$ can be easily computed. Since $\widehat{\boldsymbol{\theta}}$ is the solution to an estimating equation and $\widehat{\boldsymbol{\Sigma}}(\boldsymbol{\theta}_0)$ is a consistent estimator of $\boldsymbol{\Sigma}$ (see Section 7), the usual "*sandwich*" variance estimator of $\widehat{\boldsymbol{\theta}}$ based on Taylor's theorem is

$$\widehat{\mathbb{V}}(\widehat{\boldsymbol{\theta}}) := \Big\{\frac{1}{N}\frac{\partial \widehat{\boldsymbol{\Gamma}}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}\Big\}^{-1} \widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\theta}}) \Big\{\frac{1}{N}\frac{\partial \widehat{\boldsymbol{\Gamma}}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}\Big\}^{-1\top}. \tag{31}$$

This estimator resembles the variance estimator used with pseudo-likelihood (see Appendix C in the supplement), but there is a crucial difference. Here, we use $\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\theta}})$ instead of a two-stage variance used with pseudo-likelihood (for more details see Appendix C in the supplement). The matrix $\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\theta}})$ can be viewed as a between PSUs stratified variance estimator (Hansen & Hurwitz, 1943) which over-estimate the variance, when the sampling fraction is larger. Nevertheless, it turns out that

$\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\theta}})$ over-estimates the design variance by an amount which estimates the model variance component (see the proofs of Theorem 2 in Appendix D).

7. Main result: proof of the asymptotic consistency of the variance (25)

This Section contains the main contribution of this paper. For (29) to hold under (28), $\widehat{\boldsymbol{\Sigma}}(\boldsymbol{\theta}_0)$ needs to be a consistent estimator of the variance of $N^{-1}\widehat{\boldsymbol{\Gamma}}(\boldsymbol{\theta}_0)$. In this Section, we present the key assumptions needed for the asymptotic unbiasedness and consistency of $\widehat{\boldsymbol{\Sigma}}(\boldsymbol{\theta}_0)$. Additional regularity conditions are given in Appendix A. The proofs can be found in Appendix D of the supplement.

Using the Taylor theorem, we obtain the asymptotic unconditional "*model-design variance*" (32) of $N^{-1}\widehat{\boldsymbol{\Gamma}}(\boldsymbol{\theta}_0)$ by replacing $\widehat{\boldsymbol{B}}(\boldsymbol{\theta}_0)$ by $\boldsymbol{\beta} := \mathbb{E}_{\mathcal{M}}\mathbb{E}_d(\widehat{\boldsymbol{B}}(\boldsymbol{\theta}_0))$ within (24) (e.g. Fuller, 2009 §2.2.2).

$$\boldsymbol{\Sigma} := \frac{1}{N^2}\Big[\mathbb{E}_{\mathcal{M}}\mathbb{V}_d\big\{\widetilde{\boldsymbol{\Gamma}}(\boldsymbol{\theta}_0)\big\} + \mathbb{V}_{\mathcal{M}}\mathbb{E}_d\big\{\widetilde{\boldsymbol{\Gamma}}(\boldsymbol{\theta}_0)\big\}\Big], \tag{32}$$

where

$$\widetilde{\boldsymbol{\Gamma}}(\boldsymbol{\theta}_0) := \sum_{i\in S}\frac{1}{\pi_i}\widehat{\boldsymbol{\rho}}_i(\boldsymbol{\theta}_0) - \boldsymbol{\beta}^\top \sum_{i\in S}\frac{1}{\pi_i}\widehat{\boldsymbol{\phi}}_i\cdot \tag{33}$$

The expectation and variance under the model are denoted by $\mathbb{E}_{\mathcal{M}}$ and $\mathbb{V}_{\mathcal{M}}$. Here, $\mathbb{E}_d$ and $\mathbb{V}_d$ are the expectation and variance with respect to the design. The variance $\mathbb{V}_d\{\widetilde{\boldsymbol{\Gamma}}(\boldsymbol{\theta}_0)\}$ is the following two-stage variance

$$\mathbb{V}_d\big\{\widetilde{\boldsymbol{\Gamma}}(\boldsymbol{\theta}_0)\big\} = \mathbb{V}_d^{(1)}\Big[\mathbb{E}_d^{(2)}\{\widetilde{\boldsymbol{\Gamma}}(\boldsymbol{\theta}_0)\}\Big] + \mathbb{E}_d^{(1)}\Big[\mathbb{V}_d^{(2)}\{\widetilde{\boldsymbol{\Gamma}}(\boldsymbol{\theta}_0)\}\Big]\cdot \tag{34}$$

The operators $\mathbb{E}_d^{(1)}$ and $\mathbb{E}_d^{(2)}$ are the first and second stage expectations. The first stage refers to the selection of PSUs. The second stage is the selection of units within PSUs, which may contain more than one stage. The first and second stage variances operators are $\mathbb{V}_d^{(1)}$ and $\mathbb{V}_d^{(2)}$. The operators $\mathbb{E}_d^{(2)}$ and $\mathbb{V}_d^{(2)}$ are conditional expectations and variances given the first stage sample S. To simplify the notation, we shall use $\mathbb{E}_d^{(2)}(\cdot)$ and $\mathbb{V}_d^{(2)}(\cdot)$ instead of $\mathbb{E}_d^{(2)}(\cdot \mid S)$ and $\mathbb{V}_d^{(2)}(\cdot \mid S)$.

Since $n \to \infty$, we need an asymptotic expressions for the PSU-level joint-inclusion probabilities. We consider that the joint inclusion probabilities π_{ik} between PSUs i and k are given by the asymptotic expression of Hájek (1964, p1511):

$$\pi_{ik} := \pi_i\,\pi_k\Big\{1 - (1+\varepsilon_\pi)(1-\pi_i)(1-\pi_k)\sum_{h=1}^{H} z_{ih}\,z_{kh}\,\varphi_h^{-1}\Big\},\ \text{with } i \neq k, \tag{35}$$

where $\pi_{ik} = \pi_i$ when $i = k$, and $\varepsilon_\pi \to 0$ uniformly as $\varphi_h := \sum_{i=1}^{N} z_{ih}\,\pi_i(1 -$

$\pi_i) \to \infty$. Note that $\pi_{ik} = \pi_i \pi_k$, when i and k belongs to different strata, because $z_{ih} z_{kh} = 0$, $\forall h$. Equation (35) can be justified under under weak conditions given by Hájek (1964). Regularity conditions can be found in Berger (1998, 2011). Note that $\varphi_h \to \infty$ implies that $n_h \to \infty$ and $N_h - n_h \to \infty$, because of Chebyshev's sum inequality. Equation (35) has the advantage of allowing large sampling fraction n_h/N_h. There are enough evidences which shows that (35) is suitable for common designs (e.g. Berger, 1998, 2011; Haziza *et al.*, 2008; Matei & Tillé, 2005). It turns out that the variance (25) implicitly includes an Horvitz & Thompson (1952) variance based on (35) (for more details, see proof of Lemma 2 in the supplement).

We do not need the exact joint-inclusion probabilities for given n_h and N_h, because this would involve a non-asymptotic setup where n does not tend to ∞. Since $n \to \infty$, we have to use an asymptotic expression for these probabilities. Exact joint-inclusion probabilities would be useless to derive asymptotic properties. Only an asymptotic expression is needed. Furthermore, exact joint-inclusion probabilities relies on the π_i which are unknown.

In order to derive the asymptotic expectation of $\widehat{\boldsymbol{\Sigma}}(\boldsymbol{\theta}_0)$, we need an asymptotic expression for the PSU-level "*anticipated variance*" of (33), under informative sampling. This expression is given by the following Lemma. Its proof can be found in Appendix D.

Lemma 2 *Under* (3)*,* (4)*,* (9)*,* (10) *and* (35)*, and under the conditions* (A.1)*,* (A.2)*,* (A.3) *and* (A.12)–(A.22) *from Appendix A, we have*

$$\frac{n}{N^2}\mathbb{E}_{\mathcal{M}}\mathbb{V}_d^{(1)}\Big[\mathbb{E}_d^{(2)}\{\widetilde{\boldsymbol{\Gamma}}(\boldsymbol{\theta}_0)\}\Big] = \frac{n}{N^2}\mathbb{E}_{\mathcal{M}}\left(\sum_{i\in\mathcal{U}}\frac{1-\pi_i}{\pi_i}\boldsymbol{\epsilon}_i\,\boldsymbol{\epsilon}_i^\top\right) + o(1), \tag{36}$$

where

$$\begin{aligned}
\boldsymbol{\epsilon}_i &:= \boldsymbol{\rho}_i(\boldsymbol{\theta}_0) - \boldsymbol{\beta}^\top\boldsymbol{\phi}_i, && (37)\\
\boldsymbol{\rho}_i(\boldsymbol{\theta}_0) &:= \mathbb{E}_d^{(2)}\big\{\widehat{\boldsymbol{\rho}}_{\pi,i}(\boldsymbol{\theta}_0)\big\}, && (38)\\
\boldsymbol{\phi}_i &:= \mathbb{E}_d^{(2)}\big(\widehat{\boldsymbol{\phi}}_{\pi,i}\big), && (39)\\
\widehat{\boldsymbol{\phi}}_{\pi,i} &:= \sum_{j\in S_i}\frac{1}{\pi_{j|i}}\mathbf{f}_{j|i} && (40)
\end{aligned}$$

and $\widehat{\boldsymbol{\rho}}_{\pi,i}(\boldsymbol{\theta}_0)$ *is given by* (16)*.*

Note that under non-informative sampling, we have that (5), (A.27) and (A.28) imply that (e.g. Särndal *et al.*, 1992, p451)

$$\mathbb{E}_{\mathcal{M}}\mathbb{V}_d^{(1)}\Big[\mathbb{E}_d^{(2)}\{\widetilde{\boldsymbol{\Gamma}}(\boldsymbol{\theta}_0)\}\Big] = \sum_{i\in\mathcal{U}}\frac{1-\pi_i}{\pi_i}\mathbb{E}_{\mathcal{M}}\big(\boldsymbol{\epsilon}_i\,\boldsymbol{\epsilon}_i^\top\big)\cdot \tag{41}$$

Thus, under non-informative sampling, we see that (41) implies (36), because $(1-\pi_i)\pi_i^{-1}\mathbb{E}_{\mathcal{M}}(\boldsymbol{\epsilon}_i\,\boldsymbol{\epsilon}_i^\top) = \mathbb{E}_{\mathcal{M}}\{(1-\pi_i)\pi_i^{-1}\boldsymbol{\epsilon}_i\,\boldsymbol{\epsilon}_i^\top\}$. However, under informative sampling, (41) does not hold and $\mathbb{E}_{\mathcal{M}}\{(1-\pi_i)\pi_i^{-1}\boldsymbol{\epsilon}_i\,\boldsymbol{\epsilon}_i^\top\} \neq (1-\pi_i)\pi_i^{-1}\mathbb{E}_{\mathcal{M}}(\boldsymbol{\epsilon}_i\,\boldsymbol{\epsilon}_i^\top)$. The asymptotic expression (35) has the advantage of allowing for informative sampling and can be used for establishing (36).

Lemma 2 is key to proof Theorem 2, which establishes the asymptotic unbiasedness with (42) and consistency of $\widehat{\boldsymbol{\Sigma}}(\boldsymbol{\theta}_0)$ in (43), even though (34) is a two-stage variance.

Theorem 2 *Under the conditions of Lemmas 2 and the conditions* (A.10)*,* (A.11)*,* (A.25) *and* (A.26) *from Appendix A, we have*

$$n\mathbb{E}_{\mathcal{M}}\mathbb{E}_d\big\{\widehat{\boldsymbol{\Sigma}}(\boldsymbol{\theta}_0)\big\} = n\boldsymbol{\Sigma} + o(1), \tag{42}$$

$$n\big\{\widehat{\boldsymbol{\Sigma}}(\boldsymbol{\theta}_0) - \boldsymbol{\Sigma}\big\} = o(1)\cdot \tag{43}$$

The proofs of Theorem 2 can be found in Appendix D. Theorem 2 and Slutsky's Theorem imply (29).

8. Simulation study

We show that the proposed approach have similar performances compared to the customary pseudo-likelihood approach based on the full information about the design, given by the stratification/clustering variables, inclusion probabilities and auxiliary variables used for deriving the weights (6). The proposed empirical likelihood based inference has the advantage on not relying on these probabilities and auxiliary variables. The pseudo-likelihood approach is described in Appendix C of the supplement. We also consider a parametric model-based approach with a random intercept. The aim of the simulation studies is not to show that the proposed approach is more accurate than pseudo-likelihood. We use pseudo-likelihood as the benchmark. We want to show that the proposed approach, which does not involve the inclusion probabilities and auxiliary variables, is as accurate and has similar confidence intervals coverages as pseudo-likelihood.

We consider randomized stratified systematic samples of $n = 400$ PSUs selected from populations of different sizes containing $N = 800$, 1333, 4000 and 8000 PSUs, in order to obtain different sampling fractions: $n/N = 0.05$, 0.10, 0.30 and 0.50. The PSUs are randomly allocated to three equal sized strata. Equal allocation is used. The number N_i of units within PSU $\mathcal{U}_i$ are generated randomly using a positively skewed beta-distribution, i.e. $N_i \sim \lfloor \mathrm{Beta}(1,20)\times 280 + 20 \rfloor$. This gives $20 \leqslant N_i \leqslant 300$. The PSU-level inclusion probabilities π_i are proportional to N_i.

Within a PSU $\mathcal{U}_i$ selected, a simple random sample of size $n_i = \min\{5, \lfloor N_i/5 \rfloor\}$ is selected.

We generate covariates with symmetric, positively and negatively skewed distributions as well as one dichotomous covariate, i.e. $x_{1j} \sim (\chi^2_{df=4} - 4)8^{-\frac{1}{2}}$ a standardised χ^2-distribution, $x_{2j} \sim N(0,1)$, $x_{3j} \sim \text{Bern}(0.1)$ a Bernoulli distribution, $x_{4j} \sim \{\text{Beta}(6,2) - 0.75\}0.0208^{-\frac{1}{2}}$ a standardised beta-distribution. The response variable y_j is given by the following multiple regression model:

$$y_j = \boldsymbol{x}_j^\top \boldsymbol{\theta}_0 + \epsilon_j, \qquad j = 1, \ldots, M, \tag{44}$$

where $\epsilon_j \sim N(0,1)$, $\boldsymbol{x}_j := (1,\, x_{1j},\, x_{2j},\, x_{3j},\, x_{4j})^\top$ and $\boldsymbol{\theta}_0 = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)^\top = (1,1,1,1,1)^\top$. The vector $\boldsymbol{\theta}_0$ is the parameter of interest, i.e. $\boldsymbol{g}(\boldsymbol{Y}_j, \boldsymbol{\theta}) = \boldsymbol{x}_j(y_j - \boldsymbol{x}_j^\top \boldsymbol{\theta})$. Here, the variables y_j and $\boldsymbol{x}_j^\top$ are included within $\boldsymbol{Y}_j$

In order to control the intra-PSU correlations, a unit j with values $(y_j,\, \boldsymbol{x}_j^\top)^\top$, will be allocated randomly to a PSU in the following manner. Consider the variable $\widetilde{y}_j = y_j + e_{0j}$, where $e_{0j} \sim N(0, \text{sd} = \upsilon)$ and $j = 1, \ldots, M$. Let $\widetilde{y}_{(1)}, \ldots, \widetilde{y}_{(j)}, \ldots, \widetilde{y}_{(M)}$ be the order statistics. We allocate the unit j to the PSU $\mathcal{U}_i$, where i is such that $\widetilde{y}_{(a_{i-1})} \leqslant y_j < \widetilde{y}_{(a_i)}$. Here, $a_i := \sum_{k=1}^{i} N_k$ and $a_0 := 1$. The quantity υ controls the intra-PSU correlation. We consider $\upsilon = 3,\, 5.5,\, 8$ and 14.03 leading respectively to intra-PSU correlations of 0.3, 0.12, 0.05 and 0, i.e. the resulting design effects are approximately 10.7, 4.2, 2.6 and 1.

The survey weights $w_{j|i}$ are based on six auxiliary variables. Three auxiliary variables correlated with y_j are generated: $\xi_{1j} = (y_j - \bar{y})\sigma_y^{-1} + e_{1j}$, $\xi_{2j} = (y_j - \bar{y})\sigma_y^{-1} + e_{2j}$ and $\xi_{3j} = (y_j - \bar{y})\sigma_y^{-1} + e_{3j}$, where $e_{1j} \sim N(0, \text{sd} = 1.7)$, $e_{2j} \sim N(0, \text{sd} = 0.75)$ and $e_{3j} \sim 7 \times \{\text{Beta}(3,1) - 0.75\}$. Here, $\bar{y} := M^{-1}\sum_{j=1}^{M} y_j$ and $\sigma_y^2 := (M-1)^{-1}\sum_{j=1}^{M}(y_j - \bar{y})^2$. The correlation are $\text{corr}(\xi_{1j}, y_j) = 0.5$, $\text{corr}(\xi_{2j}, y_j) = 0.8$ and $\text{corr}(\xi_{3j}, y_j) = 0.6$. Two dichotomous auxiliary variables are generated independently according to Bernoulli distributions: $\xi_{4i} \sim \text{Bern}(0.2) - 0.2$ and $\xi_{5j} \sim \text{Bern}(0.5) - 0.5$. The last auxiliary variable is a constant variable $\xi_{6j} = M^{-1} - 1$, where M denotes the population size. The expected values of all the auxiliary variables is zero. The calibrated weights $w_{j|i}$ are derived using the "χ^2 *distance*", as in Deville & Särndal (1992).

The known parameter $\boldsymbol{d}_0$ contains the proportions of sub-groups, created by sorting the unit-level data according to the variable $\widetilde{x}_{2j} = x_{2j} + \widetilde{\epsilon}_j$ correlated with x_{2j}, where $\widetilde{\epsilon}_j \sim N(0,1)$ and $\text{corr}(x_{2j}, \widetilde{x}_{2j}) = 0.7$. The first 30% of units are allocated to the first group, the next 20% are in the second group and the remaining units belongs to the third group. The known parameter $\boldsymbol{d}_0$ is given by the proportions of units within the first two groups, i.e. $\boldsymbol{d}_0 := (0.3, 0.2)^\top$.

Thus, $\mathbf{f}(\boldsymbol{Y}_j, \boldsymbol{d}_0) = \{\delta(j \in \text{group } 1) - 0.3, \delta(j \in \text{group } 2) - 0.2\}^\top$, where $\delta(j \in \text{group } g) = 1$ if the unit j belongs to group g, and $\delta(j \in \text{group } g) = 0$ otherwise. The proportion of the third group is redundant and should not be included within $\boldsymbol{d}_0$ and $\mathbf{f}(\boldsymbol{Y}_j, \boldsymbol{d}_0)$. Here, the variables $\delta(j \in \text{group } 1)$ and $\delta(j \in \text{group } 2)$ are included within $\boldsymbol{Y}_j$

In Table 1, we have the relative efficiency defined by the MSE of the empirical likelihood estimator (17) divided by the MSE of the pseudo-likelihood estimator. We noticed a smaller MSE for the empirical likelihood estimator of the intercept. This difference is more pronounced with larger intra-PSU correlation. For the other estimators, there are negligible differences between the MSE of empirical likelihood and pseudo-likelihood.

[TABLE 1]

The observed relative biases (RB) are given in Table 2. The empirical likelihood and pseudo-likelihood approach gives similar biases. Not surprisingly, with large intra-PSU correlation, the estimators based on a random effect model can be biased. When the intra-PSU is zero, we observe no significant differences between the RB of the estimators based on empirical likelihood, pseudo-likelihood, and on a random effect model.

[TABLE 2]

In Table 3, we have the observed coverages of 95% confidence intervals. We consider two empirical likelihood confidence intervals: Wilks-type interval (30) (EL) and the Wald-type interval (ELW) based on the variance estimator (31). We also consider the pseudo-likelihood interval (PL) derived from the variance estimator which can be found in Appendix C of the supplement. We observed similar coverages for EL and ELW. For the intercept, pseudo-likelihood gives significantly low coverages decreasing with large intra-PSU correlation. The EL intervals can have better coverages than Wald-type intervals (ELW or PL). For example, with an intra-PSU correlation 0.05 and a sampling fraction 0.5, we observe low coverages (89.0% and 89.4%) for the wald-type intervals of the coefficient of the negatively skewed variable x_4. For EL the observed coverage is 95.1%. For the coefficient of the variable x_4, the coverages of ELW tend to be slightly lower than those of EL (see the last row of Table 3 which contains the column means).

[TABLE 3]

In Table 4, we have the relative biases (RB) of the empirical likelihood variance estimator (columns EL) given by (31) and the pseudo-likelihood variance estimator (columns PL) defined in Appendix C of the supplement. For the intercept, the RB of PL can be larger than the RB of EL. For the other parameters, similar RB are observed for EL and PL. The EL variance estimator is less biased, but its RB can

still be larger than 10%.

[TABLE 4]

Now, we investigate the effect of side informations (2) on the efficiency. Suppose that we use a different side information given by the means of y_j and $\widetilde{x}_{3j} = y_j + \widetilde{\epsilon}_{3i}$ correlated with y_j, where $\widetilde{\epsilon}_{3i} \sim N(0, 2)$. Thus, $\mathbf{f}(\boldsymbol{Y}_j, \boldsymbol{d}_0) = (y_j - \mu_y, \widetilde{x}_{3j} - \mu_y)^\top$, where μ_y is the expectation of the variable y_j. In Table 5, we have the observed relative efficiencies of the intercept defined as the ratio of the MSE of the EL estimator of β_0 divided by the MSE without $\mathbf{f}(\boldsymbol{Y}_j, \boldsymbol{d}_0)$, within (18). We notice a gain in efficiency with a small intra PSU correlation and a large sampling fraction. We only report the relative efficiency of the intercept, because $\mathbf{f}(\boldsymbol{Y}_j, \boldsymbol{d}_0)$ did not affect the efficiency of the estimators of β_1, β_2, β_3 and β_4, because $\mathbf{f}(\boldsymbol{Y}_j, \boldsymbol{d}_0)$ is not correlated with the residuals.

[TABLE 5]

Consider the logistic model

$$\text{logit}(y_j) = \boldsymbol{x}_j^\top \boldsymbol{\theta}_0, \qquad j = 1, \ldots, M, \tag{45}$$

where $y_j \sim \text{Bern}(P_j)$, $P_j = \text{expit}(\boldsymbol{x}_j^\top \boldsymbol{\theta}_0)$, $\boldsymbol{x}_j := (1, x_{1j}, x_{2j}, x_{3j}, x_{4j})^\top$ and $\boldsymbol{\theta}_0 = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)^\top = (-2, 1, 1, 1, 1)^\top$. The covariates follows the same distributions as in (44). The function $\boldsymbol{g}(\boldsymbol{Y}_j, \boldsymbol{\theta})$ is the estimating function of a logistic model, i.e. $\boldsymbol{g}(\boldsymbol{Y}_j, \boldsymbol{\theta}) = \boldsymbol{x}_j\{y_j - \text{expit}(\boldsymbol{x}_j^\top \boldsymbol{\theta}_0)\}$.

The auxiliary variables considered are: $\xi_{1j} = (P_j - \bar{P})\sigma_P^{-1} + e_{1j}$, $\xi_{2j} = P_j + e_{2j}$, $\xi_{3j} = P_j + e_{3j}$, $\xi_{4i} \sim \text{Bern}(0.2) - 0.2$ $\xi_{5j} \sim \text{Bern}(0.5) - 0.5$, and $\xi_{6j} = M^{-1} - 1$; where $e_{1j} \sim N(0, \text{sd} = 1.7)$, $e_{2j} \sim N(0, \text{sd} = 0.75)$ and $e_{3j} \sim 7 \times \{\text{Beta}(3, 1) - 0.75\}$. Here, $\bar{P} := M^{-1} \sum_{j=1}^{M} P_j$ and $\sigma_P^2 := (M - 1)^{-1} \sum_{j=1}^{M} (P_j - \bar{P})^2$. For the side information, we use $\mathbf{f}(\boldsymbol{Y}_j, \boldsymbol{d}_0) = \{\delta(j \in \text{group } 1) - 0.3, \delta(j \in \text{group } 2) - 0.2\}^\top$, where $\boldsymbol{d}_0 := (0.3, 0.2)^\top$ are the proportions of sub-groups, created by sorting the unit-level data according to the variable $\widetilde{x}_{2i} = x_{2i} + \widetilde{\epsilon}_i$ correlated with x_{2i}, where $\widetilde{\epsilon}_i \sim N(0, 1)$.

In Table 6, we report the observed coverages of the regression coefficients of the logistic model (45) for the proposed approach (EL) and the coverages of the Wald-type interval based on pseudo-likelihood (PL) based on the full information about the design and auxiliary variables. We observe similar coverages, although EL gives slightly smaller coverages than PL.

[TABLE 6]

9. An example of a real data application: UK PISA survey (2006)

We use the 2006 Programme for International Student Assessment (PISA) survey data for the United Kingdom. We consider a linear regression model, where '*mathematics achievement score on average*' is the response and the covariates are

City: 1 for city located schools, 0 otherwise
Parent-tertiary: 1 if parents have tertiary education, 0 otherwise
Large-class: 1 for class size over 25, 0 otherwise
Male: 1 for males and 0 for female

The data contain 13 152 students clustered into 502 schools. The school-level sampling fraction is non-negligible and approximately equal to 10%. There are nine strata. The dataset contains weights adjusted for the design and unknown auxiliary information. The inclusion probabilities and auxiliary information are not provided. Nevertheless, the minimal information is available for empirical likelihood: PSUs (schools), stratification and weights.

Since we have missing observations for some covariates, it is sensible to adjust the weights with additional variables which may explain non-response. The adjusted weights considered are the students' level weights provided in the dataset divided by fitted response probabilities computed from a logistic model with the following dichotomous covariates: *Male* (1 for males and 0 for females), *Scotland* (1 for students in Scottish schools and 0 otherwise) and *Public* (1 for students in public school and 0 otherwise). The resulting weights are the $w_{j|i}$ that are used within (7). The dataset considered contains the units with no missing values for *City*, *Parent-tertiary* and *Large-class*. The empirical likelihood approach is valid under a unit-level non-response mechanism, because we have a mutli-stage design, and the weights are adjusted for non-response at unit level.

[TABLE 7]

We consider three approaches: the empirical likelihood (EL) approach of Section 6, a parametric approach with a random effect (RE) for the intercept and Ordinary Least-Squares (OLS). Pseudo-likelihood used in Section 8 cannot be used here, because we do not know the inclusion probabilities. OLS is expected to be unreliable, because it does not take the design into account. The random effect takes the clustering into account, by allowing the intercept to vary across schools, but ignores the weights. It also relies on parametric assumption (normality of the error term and random effect) which may not hold. The estimates can also be biased, as shown in Table 2.

The results are given in Table 7. The three approaches give very different point estimates, standard deviations, p-values and confidence intervals. EL should have

the largest and more reliable standard deviations and confidence intervals, because it takes the design into account. Standard deviations should be under-estimated with the other methods. The variable *City* is significant with OLS, because of its under-estimated standard deviation. However, it is not significant with EL and RE. The variable *Large-class* is significant with EL and OLS, but not significant with RE. The variable *Male* is significant in all cases, but the confidence interval of EL is shifted upwards and only just overlap with the other intervals. EL tends to give wider confidence intervals, because it accommodates the randomness of the design and the model. Some EL confidence intervals do not overlap with OLS intervals and barely overlap with RE intervals.

This brief example shows that we may fit very different models, if we use empirical likelihood rather than random effect models. Empirical likelihood is more reliable and reveals that *Large-class* is significant. With a random effect model, this variables is not significant.

10. Discussion

The proposed approach has the advantage of only requiring calibration weights and variables specifying the stratification and identifying the PSUs. We do not rely on inclusion probabilities and auxiliary variables, which are usually unknown with public-use survey data. The PSU-level profile empirical log-likelihood test statistics follows a χ^2-distribution asymptotically under the null, even when the sampling fraction is large. It does not need a Rao & Scott (1981) adjustment or correction factors based on bootstrap, eigenvalues, design effects or finite population corrections. Therefore, it can be used like a standard likelihood ratio for testing, model building and confidence intervals. At the end of Section 7, we propose a variance estimator which does not rely on re-sampling, linearisation, inclusion probabilities or negligible sampling fractions. Note that the proposed approach relies on the condition (13), which is reasonable when fitting models on survey data.

The empirical log-likelihood ratio function implicitly recovers some asymptotic joint-inclusion probabilities between the PSUs. Indeed, this function can be approximated by a quadratic form with a consistent variance estimator which implicitly contains the asymptotic joint-inclusion probabilities (35) of Hájek (1964) within an Horvitz & Thompson (1952) variance. These probabilities do not need to be known, computed or be part of any adjustments or correction factors.

With non-informative sampling, the sampling design can be ignored under conditional model-based parametric approaches, but with informative sampling, adjustments involving inclusion probabilities are often needed. These probabilities

are not needed for the proposed unconditional approach, even under informative sampling with large sampling fractions. Calibrated weights, stratification and clustering is the only information required.

The model is specified with moment conditions without the need of a distribution for the error term. This allows unknown heteroscedasticity, because assumptions about the residuals variance are not necessary. This is an advantage over model-based parametric likelihood, with skewed data. Likelihood approaches may not be necessarily robust when the assumptions about the distribution of error terms do not hold.

The simulation study shows that the empirical likelihood approach based on minimal information (calibrated weights, stratification and clustering) gives similar (and sometime more efficient) point estimators, confidence intervals and variance estimates, compared to pseudo-likelihood, based on full information about the design, given by joint inclusion probabilities, auxiliary variables, calibrated weights, stratification and clustering. Pseudo-likelihood testing is based on Wald-type tests which can be less powerful than tests based on an empirical log-likelihood ratio function.

Optional side information can be used for empirical likelihood inference. This information is given by some variables with descriptive statistics known without error. It can be combined with empirical likelihood to improve the accuracy of estimates. Side information should not be confused with the auxiliary information used for deriving the calibrated weights. Nevertheless, it is recommended to use side information which may be correlated with some of the auxiliary variables.

Unit-level non-response can be taken into account, when the weights have been adjusted for non-response, under the usual "*missing at random*" response mechanism. In this case, non-response is in-bedded within the design as an ultimate stage. The non-response variables used for adjusting the calibration weights are not needed. However, there are some limitations. If non-response occurs at PSU-level or we have a single-stage design, the point estimator is still consistent, but the property (29) would not hold, because we assumed that a fixed number of PSUs is selected. In this case, we would need to know the variables used for non-response weighting, because they would need to be part of the side information as in Berger (2018).

Supplementary material

The online supplementary material contains Appendix B, C and D. Algorithms for the computation of the point estimator and the empirical log-likelihood ratio

function are given in Appendix B. The Pseudo-likelihood approach implemented in Section 8 is described in Appendix C. Appendix D contains the proofs.

References

Berger, Y. G. (1998) Rate of convergence to asymptotic variance for the Horvitz-Thompson estimator. *J. Statist. Plann. Inference*, **74**, 149–168.

Berger, Y. G. (2011) Asymptotic consistency under large entropy sampling designs with unequal probabilities. *Pakistan Journal of Statistics*, **27**, 407–426.

Berger, Y. G. (2018) An empirical likelihood approach under cluster sampling with missing observations. *Ann. Inst. Stat. Math.*, **72**, 91–121.

Berger, Y. G. & Patilea, V. (2022) A semi-parametric empirical likelihood approach for conditional estimating equations under endogenous selection. *Econom. Stat.*

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*, 15. URL `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. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, **78**, 319–341.

Bertail, P., Chautru, E., & Clémencon, S. (2017) Empirical processes in survey sampling with (conditional)poisson designs. *Scand. J. Stat.*, **44**, 97–111.

Besag, J. (1975) Statistical analysis of non-lattice data. *J. R. Stat. Soc. Ser. C. The Statistician*, **24**, 179–195.

Binder, D. A. & Roberts, G. (2009) Design- and model-based inference for model parameters. In *Sample Surveys: Design, Methods and Applications* (eds. D. Pfeffermann & C. R. Rao), vol. 29B of *Handbook of Statistics*, 33–54. Amsterdam: Elsevier.

Chambers, R. L. & Skinner, C. J., eds. (2003) *Analysis of Survey Data*. Chichester: Wiley.

Chambers, R. L., Steel, D. G., Wang, S. & Welsh, A. (2012) *Maximum Likelihood Estimation for Sample Surveys*. New York: Chapman and Hall/CRC.

Chaudhuri, S. & Handcock, M. S. (2018) A conditional empirical likelihood based

method for model parameter estimation from complex survey datasets. *Statistics and Applications*, **16**, 245–268.

Chaudhuri, S., Handcock, M. S. & Rendall, M. S. (2008) Generalized linear models incorporating population level information: An empirical-likelihood-based approach. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, **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, S. & Kim, J. K. (2014) Population empirical likelihood for nonparametric inference in survey sampling. *Statist. Sinica*, **24**, 335–355.

Deville, J. C. & Särndal, C.-E. (1992) Calibration estimators in survey sampling. *J. Amer. Statist. Assoc.*, **87**, 376–382.

Domínguez, M. A. & Lobato, I. N. (2004) Consistent estimation of models defind by conditional moment restrictions. *Econometrika*, **72**, 1601–1615.

Donald, S., Imbens, G. & Newey, W. (2009) Choosing instrumental variables in conditional moment restriction models. *J. Econometrics*, **152**, 28–36.

Eurostat (2012) European union statistics on income and living conditions (EU-SILC). `http://epp.eurostat.ec.europa.eu/portal/page/portal/microdata/eu_silc`.

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

Hájek, J. (1964) Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Ann. Math. Stat.*, **35**, 1491–1523.

Hansen, M. H. & Hurwitz, W. N. (1943) On the theory of sampling from finite populations. *Ann. Math. Stat.*, **14**, pp. 333–362.

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

Horvitz, D. G. & Thompson, D. J. (1952) A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.*, **47**, 663–685.

Imbens, G. W. & Lancaster, T. (1994) Combining micro and macro data in microeconometric models. *Rev. Econ. Stud.*, **61**, 655–680.

Isaki, C. T. & Fuller, W. A. (1982) Survey design under the regression superpopulation model. *J. Amer. Statist. Assoc.*, **77**, 89–96.

Kim, J. K. (2009) Calibration estimation using empirical likelihood in survey sampling. *Statist. Sinica*, **19**, 145–157.

Krieger, A. M. & Pfeffermann, D. (1992) Maximum likelihood estimation from complex sample surveys. *Survey Methodology*, **18**, 225–239.

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.

National Center for Health Statistics (2016) National health and nutrition examination survey (NHANES). `http://www.cdc.gov/nchs/nhanes`.

Newey, W. K. (1993) Efficient estimation of models with conditional moment restrictions. In *Econometrics* (eds. G. Maddala, C. Rao & H. Vinod), vol. 11 of *Handbook of Statistics*, 2111–2245. Amsterdam: Elsevier.

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

Oğuz-Alper, M. & Berger, Y. G. (2016b) Online supplementary materials of Oğuz-Alper & Berger (2016a). *Biometrika*, 10pp.

Pfeffermann, D., Krieger, A. & Rinott, Y. (1998) Parametric distributions of complex survey data under informative probability sampling. *Statist. Sinica*, **8**, 1087–1114.

Pfeffermann, D. & Sverchkov, M. (1999) Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhya B*, **61**, 166–186.

Qin, J. & Lawless, J. (1994) Empirical likelihood and general estimating equations. *Ann. Statist.*, **22**, pp. 300–325.

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. *J. Amer. Statist. Assoc.*, **76**, 221–230.

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.

Robinson, P. M. & Särndal, C. E. (1983) Asymptotic properties of the generalized regression estimator in probability sampling. *Sankhya B*, **43**, 240–248.

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

Skinner, C., Holt, D. & Smith, T., eds. (1989) *Analysis of complex surveys*. Chichester: Wiley.

Wu, C. & Rao, J. N. K. (2006) Pseudo-empirical likelihood ratio confidence intervals for complex surveys. *Canad. J. Statist.*, **34**, 359–375.

Zhao, P., Ghosh, M., Rao, J. & Wu, C. (2019) Bayesian empirical likelihood inference with complex survey data. *J. R. Stat. Soc. Ser. B. Stat. Methodol. (Statistical Methodology)*.

Zhao, P., Rao, J. N. K. & Wu, C. (2020) Empirical likelihood inference with public-use survey data. *Electron. J. Stat.*, **14**, 2484 – 2509.

Zhao, P. & Wu, C. (2018) Some theoretical and practical aspects of empirical likelihood methods for complex surveys. *Int. Stat. Rev.*, **87**, 239–256.

Address for correspondence:
Yves G. Berger, Southampton Statistical Sciences Research Institute, University of Southampton, Highfield, Southampton, SO17 1BJ, UK.
E-mail address: *y.g.berger@soton.ac.uk*
URL: `http://www.yvesberger.co.uk`
ORCID: `http://orcid.org/0000-0002-9128-5384`

Appendix A: Regularity conditions

We assume that the following regularity conditions hold.

$$\frac{1}{N}\Big\|\sum_{i\in S}\frac{1}{\pi_i}\widehat{\boldsymbol{\phi}}_i\Big\| = O_p(n^{-\frac{1}{2}}), \tag{A.1}$$

$$\frac{1}{N}\Big\|\sum_{i\in S}\frac{1}{\pi_i}\mathbb{E}_d^{(2)}\big(\widehat{\boldsymbol{\phi}}_i\big)\Big\| = O_p(n^{-\frac{1}{2}}), \tag{A.2}$$

$$\exists \varsigma \in \mathbb{R}: \quad \mathbb{P}\big(nN^{-1}\,\pi_i^{-1} \leqslant \varsigma\big) \to 1, \qquad \forall i \in \mathcal{U}, \tag{A.3}$$

$$\mathbb{E}\big\{\|\widehat{\boldsymbol{\varsigma}}_i^*(\boldsymbol{\theta})\|^4\big\} < \infty, \qquad \forall\, \boldsymbol{\theta} \in \mathcal{B}_n, \tag{A.4}$$

$$\|\dot{\boldsymbol{\Gamma}}(\boldsymbol{\theta}_0)\| = O_p(1), \quad \text{where } \dot{\boldsymbol{\Gamma}}(\boldsymbol{\theta}) := \frac{\partial \widehat{\boldsymbol{\Gamma}}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}, \tag{A.5}$$

where $\widehat{\boldsymbol{\phi}}_i := \pi_i \breve{\mathbf{f}}_i$ and

$$\begin{aligned} \widehat{\boldsymbol{\varsigma}}_i^*(\boldsymbol{\theta}) &:= \pi_i\, \breve{\boldsymbol{c}}_i^*(\boldsymbol{\theta}), \\ \mathcal{B}_n &:= \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leqslant b_n\}\cdot \end{aligned} \tag{A.6}$$

The vector $\breve{\boldsymbol{c}}_i^*(\boldsymbol{\theta})$ is defined by (21).

Conditions (A.1)–(A.2) means that the law-of-large-numbers holds for $\widehat{\boldsymbol{\phi}}_i$ and $\mathbb{E}_d\{\widehat{\boldsymbol{\phi}}_i\}$ (e.g. Isaki & Fuller, 1982) as in (9) and (10). Condition (A.3) is a standard assumption which ensures that π_i is not disproportionately small compared to n/N (Isaki & Fuller, 1982). Condition (A.4) ensures that the fourth moments of $\widehat{\boldsymbol{\varsigma}}_i^*(\widetilde{\boldsymbol{\theta}})$ exists. Condition (A.5) assumes that the gradient is bounded for $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

We assume that there exists constants γ_S and γ_G such that

$$\mathbb{P}\Big[\inf_{\boldsymbol{\theta}\in\mathcal{B}_n} \gamma_{\min}\big\{\widehat{\boldsymbol{\Phi}}^*(\boldsymbol{\theta})\big\} \geqslant \gamma_S > 0\Big] \to 1, \quad \forall\, \boldsymbol{\theta} \in \mathcal{B}_n, \tag{A.7}$$

$$\mathbb{P}\Big[\inf_{\boldsymbol{\theta}\in\mathcal{B}_n} \gamma_{\min}\big\{\dot{\boldsymbol{\Gamma}}(\boldsymbol{\theta}_0)^\top \dot{\boldsymbol{\Gamma}}(\boldsymbol{\theta}_0)\big\} \geqslant \gamma_G > 0\Big] \to 1, \tag{A.8}$$

where $\gamma_{\min}\{\boldsymbol{A}\}$ denotes the smallest eigenvalue of $\boldsymbol{A}$ when $\boldsymbol{A}$ is symmetric. Here,

$$\widehat{\boldsymbol{\Phi}}^{*}(\boldsymbol{\theta}) := \frac{n}{N^2}\sum_{i\in S}\frac{1}{\pi_i^2}\widehat{\boldsymbol{\varsigma}}_i^{*}(\boldsymbol{\theta})\,\widehat{\boldsymbol{\varsigma}}_i^{*}(\boldsymbol{\theta})^{\top}, \tag{A.9}$$

where $\widehat{\boldsymbol{\varsigma}}_i^{*}(\boldsymbol{\theta})$ is defined by (A.6).

Condition (A.7) states that the variance-covariance matrix (A.9) is invertible within the neighbourhood $\mathcal{B}_n$. Inequality (A.8) is a mild condition which states that the Gramian matrix $\dot{\boldsymbol{\Gamma}}(\boldsymbol{\theta}_0)^{\top}\dot{\boldsymbol{\Gamma}}(\boldsymbol{\theta}_0)$ is positive definite asymptotically.

We assume that

$$\|\widehat{\boldsymbol{\Phi}}^{*}(\boldsymbol{\theta}_0) - \boldsymbol{\Phi}\| = o_p(1), \tag{A.10}$$

$$\|\widehat{\boldsymbol{\Phi}}^{*}(\boldsymbol{\theta})\| = O_p(1), \quad \forall\,\boldsymbol{\theta}\in\mathcal{B}_n, \tag{A.11}$$

$$\exists\,\mathscr{L}^{*}\in\mathbb{R}: \quad \mathbb{E}\big(\mathscr{L}^{*}\big) < \infty \quad \text{and} \quad \big\|\widehat{\boldsymbol{\Phi}}^{*}(\boldsymbol{\theta}_0)\big\| \leqslant \mathscr{L}^{*}, \tag{A.12}$$

where

$$\begin{aligned}\boldsymbol{\Phi} &:= \frac{n}{N^2}\mathbb{E}_{\mathcal{M}}\mathbb{E}_{d}\bigg\{\sum_{i\in S}\frac{1}{\pi_i^2}\widehat{\boldsymbol{\varsigma}}_{\pi,i}^{*}(\boldsymbol{\theta}_0)\,\widehat{\boldsymbol{\varsigma}}_{\pi,i}^{*}(\boldsymbol{\theta}_0)^{\top}\bigg\}, \\ \widehat{\boldsymbol{\varsigma}}_{\pi,i}^{*}(\boldsymbol{\theta}) &:= \Big\{\frac{N}{n}\pi_i\boldsymbol{z}_i^{\top}, \widehat{\boldsymbol{\phi}}_{\pi,i}^{\top}, \widehat{\boldsymbol{\rho}}_{\pi,i}(\boldsymbol{\theta})^{\top}\Big\}^{\top}.\end{aligned} \tag{A.13}$$

We also assume that $\|\boldsymbol{\Phi}\| < \infty$. Note that (A.13) is positive definite. Conditions (A.10)–(A.12) state that the variance-covariance matrix $\widehat{\boldsymbol{\Phi}}^{*}(\boldsymbol{\theta}_0)$ converges to $\boldsymbol{\Phi}$, is bounded within the neighbourhood $\mathcal{B}_n$ and dominated at $\boldsymbol{\theta}_0$.

We assumed that the side information is such that

$$\|\widehat{\boldsymbol{\Phi}}^{-1}\| = O_p(1), \tag{A.14}$$

holds, where

$$\widehat{\boldsymbol{\Phi}} := \frac{n}{N^2}\sum_{i\in S}\frac{1}{\pi_i^2}\boldsymbol{\varsigma}_i\,\boldsymbol{\varsigma}_i^{\top}, \tag{A.15}$$

$$\begin{aligned}\boldsymbol{\varsigma}_i &:= \pi_i\,\breve{\boldsymbol{c}}_i, \\ \breve{\boldsymbol{c}}_i &:= \Big(\frac{N}{n}\boldsymbol{z}_i^{\top}, \breve{\mathbf{f}}_i^{\top}\Big)^{\top}.\end{aligned} \tag{A.16}$$

It is common practice to assume (A.14) (e.g Deville & Särndal, 1992, proof of result 3, page 381). The condition (A.10) implies that $\|\widehat{\boldsymbol{\Phi}}^{*}(\boldsymbol{\theta}_0)^{-1}\| = O_p(1)$. Condition (A.14) states that the norm of the inverse of the sub-matrix (A.15) of $\widehat{\boldsymbol{\Phi}}^{*}(\boldsymbol{\theta}_0)$, is also bounded.

We assume that for $\ell = 1, \dots d_{\mathbf{f}}$

$$\frac{1}{N}\sum_{i\in S}\frac{1}{\pi_i}\widehat{\phi}_i^{(\ell)} = \frac{1}{N}\sum_{i\in S}\frac{1}{\pi_i}\widehat{\phi}_{\pi,i}^{(\ell)} + o_p(n^{-\frac{1}{2}}) \quad \text{or } \frac{1}{N}\sum_{i\in S}\frac{1}{\pi_i}\widehat{\phi}_i^{(\ell)} = o_p(n^{-\frac{1}{2}})\cdot \quad \text{(A.17)}$$

Here, $\widehat{\phi}_i^{(\ell)}$ and $\widehat{\phi}_{\pi,i}^{(\ell)}$ are respectively the ℓ-th component of $\widehat{\boldsymbol{\phi}}_i$ and $\widehat{\boldsymbol{\phi}}_{\pi,i}$, where $\widehat{\boldsymbol{\phi}}_i$ is defined by (19) and $\widehat{\boldsymbol{\phi}}_{\pi,i}$ is given by (40). Condition (A.17) is equivalent to (15), but for the side information.

In order to ensure convergence in expectation, we assume that some random variables are dominated, i.e. $\forall h = 1, \dots, H$,

$$\exists\, \mathscr{C}_h \in \mathbb{R}: \quad \mathbb{E}\big(\mathscr{C}_h\big) < \infty \quad \text{and} \quad \frac{n_h^{\frac{1}{2}}}{N_h}\Big\|\sum_{i\in S_h}\frac{1}{\pi_i}\boldsymbol{\rho}_i(\boldsymbol{\theta}_0)\Big\| \leqslant \mathscr{C}_h, \quad \text{(A.18)}$$

$$\exists\, \widehat{\mathscr{C}}_h \in \mathbb{R}: \quad \mathbb{E}\big(\widehat{\mathscr{C}}_h\big) < \infty \quad \text{and} \quad \frac{n_h^{\frac{1}{2}}}{N_h}\Big\|\sum_{i\in S_h}\frac{1}{\pi_i}\widehat{\boldsymbol{\rho}}_i(\boldsymbol{\theta}_0)\Big\| \leqslant \widehat{\mathscr{C}}_h, \quad \text{(A.19)}$$

$$\exists\, \widetilde{\mathscr{C}} \in \mathbb{R}: \quad \mathbb{E}\big(\widetilde{\mathscr{C}}\big) < \infty \quad \text{and} \quad \frac{n^{\frac{1}{2}}}{N}\Big\|\sum_{i\in S}\frac{1}{\pi_i}\widehat{\boldsymbol{\rho}}_{\pi,i}(\boldsymbol{\theta}_0)\Big\| \leqslant \widetilde{\mathscr{C}}, \quad \text{(A.20)}$$

$$\exists\, \mathscr{E}_h \in \mathbb{R}: \quad \mathbb{E}\big(\mathscr{E}_h\big) < \infty \quad \text{and} \quad \frac{n_h^{\frac{1}{2}}}{N_h}\Big\|\sum_{i\in S_h}\frac{1}{\pi_i}\boldsymbol{\phi}_i\Big\| \leqslant \mathscr{E}_h, \quad \text{(A.21)}$$

$$\exists\, \widehat{\mathscr{E}}_h \in \mathbb{R}: \quad \mathbb{E}\big(\widehat{\mathscr{E}}_h\big) < \infty \quad \text{and} \quad \frac{n_h^{\frac{1}{2}}}{N_h}\Big\|\sum_{i\in S_h}\frac{1}{\pi_i}\widehat{\boldsymbol{\phi}}_i\Big\| \leqslant \widehat{\mathscr{E}}_h, \quad \text{(A.22)}$$

$$\exists\, \widetilde{\mathscr{E}} \in \mathbb{R}: \quad \mathbb{E}\big(\widetilde{\mathscr{E}}\big) < \infty \quad \text{and} \quad \frac{n^{\frac{1}{2}}}{N}\Big\|\sum_{i\in S}\frac{1}{\pi_i}\widehat{\boldsymbol{\phi}}_{\pi,i}\Big\| \leqslant \widetilde{\mathscr{E}}, \quad \text{(A.23)}$$

$$\exists\, \mathscr{H} \in \mathbb{R}: \quad \mathbb{E}(\mathscr{H}) < \infty \quad \text{and} \quad \frac{n}{N^2}\Big\|\sum_{i\in\mathcal{U}}\frac{1}{\pi_i}\boldsymbol{\epsilon}_i\,\boldsymbol{\epsilon}_i^\top\Big\| \leqslant \mathscr{H}, \quad \text{(A.24)}$$

where $\boldsymbol{\epsilon}_i$, $\boldsymbol{\rho}_i(\boldsymbol{\theta}_0)$ and $\boldsymbol{\phi}_i$ are defined respectively by (37), (38) and (39). We also have the following regularity conditions for the regression coefficient $\widehat{\boldsymbol{B}}(\boldsymbol{\theta}_0)$.

$$\exists\, \mathscr{B} \in \mathbb{R}: \quad \mathbb{E}\big(\mathscr{B}\big) < \infty, \quad |\mathscr{B}| = O_p(1) \quad \text{and} \quad \big\|\widehat{\boldsymbol{B}}(\boldsymbol{\theta}_0)\big\| \leqslant \mathscr{B}, \quad \text{(A.25)}$$

$$\|\boldsymbol{\beta}\| < \infty \quad \text{and} \quad \|\widehat{\boldsymbol{B}}(\boldsymbol{\theta}_0) - \boldsymbol{\beta}\| = o_p(1), \quad \text{where } \boldsymbol{\beta} := \mathbb{E}_{\mathcal{M}}\mathbb{E}_d(\widehat{\boldsymbol{B}}(\boldsymbol{\theta}_0))\cdot \quad \text{(A.26)}$$

Equivalently to (5), we assume

$$\mathbf{f}_{j|i} \perp\!\!\!\perp \mathbf{f}_{\ell|k} \qquad \forall\, j \in \mathcal{U}_i,\ \ell \in \mathcal{U}_k \text{ and } i \neq k; \quad \text{(A.27)}$$

$$\boldsymbol{g}_{j|i}(\boldsymbol{\theta}_0) \perp\!\!\!\perp \mathbf{f}_{\ell|k} \qquad \forall\, j \in \mathcal{U}_i,\ \ell \in \mathcal{U}_k \text{ and } i \neq k\cdot \quad \text{(A.28)}$$

Table 1: Relative efficiencies defined as the MSE of the empirical likelihood point estimator divided by the MSE of the pseudo-likelihood estimator.

Intra PSU Correlation	n/N	β_0	β_1	β_2	β_3	β_4
0.00	0.05	0.98	1.01	1.01	1.00	1.01
	0.10	0.99	1.01	1.00	1.01	1.01
	0.30	0.99	1.00	1.01	1.02	1.01
	0.50	1.00	1.00	1.01	1.01	1.00
0.05	0.05	0.98	1.00	1.00	1.01	1.01
	0.10	0.97	1.00	1.00	1.01	1.01
	0.30	0.97	1.00	1.00	1.00	1.01
	0.50	0.99	1.00	1.00	1.01	1.01
0.12	0.05	0.97	1.01	1.00	1.00	1.00
	0.10	0.95	1.00	1.00	1.01	1.00
	0.30	0.98	1.00	1.01	1.01	1.00
	0.50	0.99	1.00	1.00	1.00	1.00
0.30	0.05	0.94	1.01	1.00	1.01	1.00
	0.10	0.91	1.00	1.01	1.00	1.00
	0.30	0.93	1.00	1.00	1.00	1.01
	0.50	0.98	1.01	0.98	1.00	1.00

Table 2: Observed Relative biases (%) of the regression coefficients. EL: empirical likelihood estimator. PEL: Pseudo-likelihood estimator. RE: Parametric estimator based on a random effect model.

Intra PSU Corr.	n/N	Intercept			$x_1 \sim (\chi^2_{df=4} - 4)/8^{\frac{1}{2}}$			$x_2 \sim N(0,1)$			$x_3 \sim \text{Bern}(0.1)$			$x_4 \sim \text{Beta}_s(6,2)^*$		
		EL	PL	RE	EL	PL	RE	EL	PL	RE	EL	PL	RE	EL	PL	RE
0.00	0.05	0.1	0.1	0.1	−0.1	−0.1	−0.1	0.0	0.0	0.0	−0.3	−0.3	−0.3	0.0	0.0	0.0
	0.10	0.1	0.1	0.1	0.0	0.0	−0.1	−0.1	−0.1	−0.1	0.0	0.0	0.1	0.0	0.0	0.0
	0.30	−0.1	−0.1	−0.1	0.1	0.1	0.0	0.0	0.0	−0.1	0.0	0.0	0.0	−0.1	−0.1	−0.1
	0.50	0.2	0.2	0.2	0.0	0.0	0.0	0.0	0.0	−0.1	0.0	0.1	0.0	0.0	0.0	0.0
0.05	0.05	−0.1	−0.1	−0.1	−0.1	−0.1	−0.2	0.0	0.0	−0.1	−0.1	−0.1	−0.2	0.0	0.0	−0.1
	0.10	0.1	0.1	0.1	0.0	0.0	−0.2	0.0	0.0	−0.1	−0.3	−0.3	−0.4	−0.1	−0.1	−0.3
	0.30	0.0	0.0	0.0	0.0	0.0	−0.2	0.0	0.0	−0.2	0.2	0.2	0.0	0.0	0.0	−0.2
	0.50	0.0	−0.2	0.0	−0.4	−0.5	−0.2	0.6	0.6	−0.1	−1.8	−1.8	−0.1	1.5	1.5	−0.2
0.12	0.05	0.0	0.0	0.1	0.0	0.0	−0.5	−0.1	−0.1	−0.6	0.1	0.1	−0.5	0.0	0.0	−0.6
	0.10	−0.1	−0.1	0.0	−0.1	−0.1	−0.6	−0.1	−0.1	−0.7	−0.1	−0.2	−0.7	0.1	0.1	−0.5
	0.30	0.2	0.2	0.2	−0.1	−0.1	−0.7	0.1	0.1	−0.4	0.1	0.1	−0.4	0.0	0.0	−0.6
	0.50	−0.1	−0.1	0.0	0.0	0.0	−0.6	0.0	0.0	−0.5	0.1	0.1	−0.4	−0.1	−0.1	−0.7
0.30	0.05	−0.1	−0.1	0.3	−0.1	−0.1	−3.8	0.0	0.0	−3.7	0.1	0.1	−3.6	0.0	0.0	−3.7
	0.10	0.0	0.0	0.3	0.0	0.0	−3.7	0.1	0.1	−3.6	0.1	0.2	−3.6	0.0	0.0	−3.6
	0.30	0.2	0.2	0.6	0.1	0.1	−3.5	0.0	0.0	−3.6	0.0	0.1	−3.5	−0.1	−0.1	−3.7
	0.50	−0.1	−0.1	0.4	−0.1	−0.1	−3.7	−0.1	−0.1	−3.7	−0.3	−0.3	−3.8	−0.1	−0.1	−3.8

* $\text{Beta}_s(6,2)$ is the standardised Beta distribution defined by $\text{Beta}_s(6,2) := \{\text{Beta}(6,2) - 0.75\}/0.0208^{\frac{1}{2}}$

Table 3: Observed coverages (%) of the regression coefficients. EL: coverages based on the empirical log-likelihood ratio function. ELW: coverages of the Wald-type statistics based on the variance estimator (31) of the empirical likelihood estimator. PL: coverages of the Wald-type statistics based on the variance estimator of the pseudo-likelihood estimator

Intra PSU Corr.	n/N	Intercept			$x_1 \sim (\chi^2_{df=4} - 4)/8^{\frac{1}{2}}$			$x_2 \sim N(0,1)$			$x_3 \sim \text{Bern}(0.1)$			$x_4 \sim \text{Beta}_s(6,2)^*$		
		EL	ELW	PL	EL	ELW	PL	EL	ELW	PL	EL	ELW	PL	EL	ELW	PL
0.00	0.05	95.0	94.8	91.3†	93.7	93.5†	94.1	95.3	95.0	95.3	95.4	95.5	95.7	95.0	94.7	95.2
	0.10	94.8	94.7	91.3†	94.9	94.8	95.1	95.1	95.2	95.3	95.6	95.4	96.0	94.1	93.8	94.5
	0.30	96.1	96.0	92.2†	94.1	93.9	93.8	93.7	93.7	94.2	95.0	95.0	95.1	94.6	94.4	94.7
	0.50	94.1	93.7	91.3†	95.3	95.2	95.2	95.1	94.8	95.1	94.0	93.7	94.2	94.4	94.4	94.6
0.05	0.05	94.7	94.7	92.0†	94.5	94.4	94.7	94.0	94.0	94.1	94.8	94.8	95.1	96.0	95.9	96.3
	0.10	94.5	94.2	90.3†	94.2	93.9	94.7	94.3	94.2	94.4	94.6	94.4	95.0	94.7	94.6	94.9
	0.30	96.2	96.2	91.9†	95.3	95.3	95.6	94.2	94.2	94.1	95.4	95.5	95.3	95.4	95.3	95.5
	0.50	95.5	96.3	93.0†	94.6	94.5	95.0	94.7	95.2	95.3	95.0	94.7	94.8	95.1	89.0†	89.4†
0.12	0.05	96.0	95.8	89.8†	94.6	94.6	94.7	93.8	93.8	93.9	94.4	94.0	94.5	94.7	94.6	94.7
	0.10	95.4	95.3	91.6†	96.7†	96.4†	96.5†	96.3	96.2	96.0	94.5	94.5	95.0	93.8	93.8	93.7
	0.30	94.8	94.8	90.4†	94.5	94.3	94.4	94.3	94.3	94.3	94.6	94.3	94.7	95.0	94.9	94.9
	0.50	96.6†	96.4†	92.8†	95.5	95.5	95.5	94.8	94.5	94.4	93.6†	93.7	93.7	95.1	94.9	95.0
0.30	0.05	94.7	94.6	87.1†	93.4†	93.3†	93.2†	94.2	94.2	94.2	93.9	93.7	93.8	94.2	94.1	94.4
	0.10	95.2	94.9	87.4†	94.5	94.5	94.5	93.7	93.6†	93.8	94.3	94.2	94.2	95.0	94.7	94.7
	0.30	95.1	94.9	88.1†	95.2	95.0	94.8	94.8	94.7	94.7	95.2	95.1	95.3	93.8	93.8	93.8
	0.50	96.6†	96.4†	89.7†	95.1	95.1	95.4	94.4	94.3	93.8	95.0	95.0	94.8	93.9	93.8	93.7
Column means:		95.3	95.2	89.6	94.8	94.7	94.9	94.8	94.7	94.7	94.7	94.6	94.8	94.7	94.2	94.3

* $\text{Beta}_s(6,2)$ is the standardised Beta distribution defined by $\text{Beta}_s(6,2) := \{\text{Beta}(6,2) - 0.75\}/0.0208^{\frac{1}{2}}$

† Coverages significantly different from 95%. P-value ≤ 0.05

Table 4: Relative biases (%) of variance estimates of the regression coefficients.

Intra PSU Corr.	n/N	$x_0 = 1$ (int.) EL	$x_0 = 1$ (int.) PL	$x_1 \sim (\chi^2_4 - 4)/8^{\frac{1}{2}}$ EL	$x_1 \sim (\chi^2_4 - 4)/8^{\frac{1}{2}}$ PL	$x_2 \sim N(0,1)$ EL	$x_2 \sim N(0,1)$ PL	$x_3 \sim \text{Bern}(0.1)$ EL	$x_3 \sim \text{Bern}(0.1)$ PL	$x_4 \sim \text{Beta}_s(6,2)^*$ EL	$x_4 \sim \text{Beta}_s(6,2)^*$ PL
0.00	0.05	−2.4	−22.1	−8.6	−7.1	−1.9	0.0	0.3	1.0	−6.5	−4.1
	0.10	−1.4	−20.2	2.9	5.0	−1.0	0.2	0.7	2.7	1.4	3.5
	0.30	0.0	−18.7	−5.1	−3.7	−8.1	−6.5	0.8	3.3	−3.1	−1.6
	0.50	−3.5	−21.2	−0.2	1.1	0.0	1.8	−4.3	−2.7	−3.8	−3.3
0.05	0.05	1.6	−21.8	−1.5	−0.3	−2.8	−2.1	−6.3	−4.8	9.7	12.3
	0.10	0.9	−23.3	0.1	1.5	0.0	1.4	−2.5	−0.7	−4.6	−2.7
	0.30	1.7	−22.6	3.6	4.3	−8.1	−6.8	−0.3	0.6	3.1	4.5
	0.50	14.3	−10.9	4.3	6.1	20.2	22.6	0.4	1.3	8.1	9.3
0.12	0.05	−2.0	−29.6	−1.2	0.9	−7.1	−5.8	−9.0	−7.6	−0.3	1.0
	0.10	4.8	−26.1	6.5	6.9	5.1	5.6	−3.8	−2.2	−4.9	−3.7
	0.30	4.5	−24.0	−0.3	0.1	−7.6	−5.8	−0.7	1.0	−2.3	−2.0
	0.50	11.4	−19.1	3.4	4.1	0.4	0.7	−5.5	−4.1	−3.4	−2.9
0.30	0.05	−3.4	−41.6	−10.4	−9.4	−2.0	−1.1	−7.4	−5.4	−5.9	−5.1
	0.10	0.1	−40.7	−0.8	−0.7	−11.6	−10.7	−2.2	−1.0	−3.8	−3.4
	0.30	6.8	−36.7	−1.0	−2.9	−1.8	−2.9	−0.5	0.7	−2.3	−2.8
	0.50	7.5	−33.6	4.3	2.1	−0.3	−4.6	−4.9	−3.9	−1.9	−4.1
Column means:		2.6	−25.8	−0.2	0.5	−1.7	−0.9	−2.8	−1.4	−1.3	−0.3

* $\text{Beta}_s(6,2)$ is the standardised Beta distribution defined by $\text{Beta}_s(6,2) := \{\text{Beta}(6,2) - 0.75\}/0.0208^{\frac{1}{2}}$

Table 5: Relative efficiencies of the intercept defined as the ratio of the MSE of the EL estimator of β_0 divided by the MSE without $\mathbf{f}(\boldsymbol{Y}_j, \boldsymbol{d}_0)$, within (18).

Intra PSU Correlation	n/N 0.05	0.1	0.3	0.5
0.00	0.77	0.74	0.75	0.75
0.05	0.70	0.73	0.72	0.79
0.12	0.67	0.64	0.74	0.74
0.30	0.53	0.56	0.62	0.67
0.50	0.51	0.52	0.53	0.60

Table 6: Observed coverages (%) of the regression coefficients of the logistic model (45). EL: coverages based on the empirical log-likelihood ratio function. PL: coverages of the Wald-type statistics based on the variance estimator of the pseudo-likelihood estimator. $x_1 \sim (\chi^2_{df=4} - 4)/8^{\frac{1}{2}}$, $x_2 \sim N(0,1)$, $x_3 \sim \text{Bern}(0.1)$, $x_4 \sim \{\text{Beta}(6,2) - 0.75\}/0.0208^{\frac{1}{2}}$

	Intercept		x_1		x_2		x_3		x_4	
n/N	EL	PL	EL	PL	EL	PL	EL	PL	EL	PL
0.05	94.8	94.9	93.2†	93.6†	94.5	94.7	94.8	94.8	94.3	94.4
0.10	94.0	94.4	93.3†	93.8	94.5	94.8	93.5†	94.2	95.0	95.0
0.30	93.8	94.3	94.0	94.3	96.0	96.3	93.6†	94.2	94.7	95.1
0.50	94.7	95.1	94.7	95.1	95.2	95.3	95.1	95.5	94.7	95.1

† Coverages significantly different from 95%. P-value ≤ 0.05

Table 7: Pisa estimates (Est.), standard deviations (SD), p-values and bounds of 95% confidence intervals (Conf. Int.) for three methods: Empirical likelihood (EL), parametric approach with a Random Effect (RE) and Ordinary Least-Squares (OLS).

					95% Conf. Int.	
Covariates		Est.	SD	P-values	Lower	Upper
Constant	EL	493.3	5.7	0.000	481.5	505.2
	RE	488.1	3.4	0.000	481.5	494.8
	OLS	485.6	1.6	0.000	482.6	488.7
City	EL	−7.0	8.8	0.477	−25.7	9.8
	RE	−8.6	4.9	0.075	−18.2	0.9
	OLS	−4.8	1.7	0.006	−8.2	−1.4
Parent-tertiary	EL	29.0	3.7	0.000	21.8	36.5
	RE	11.5	1.4	0.000	8.8	14.1
	OLS	24.5	1.5	0.000	21.5	27.5
Large-class	EL	−20.3	6.3	0.001	−32.8	−7.9
	RE	−1.3	4.4	0.763	−10.0	7.4
	OLS	−3.3	1.6	0.034	−6.4	−0.3
Male	EL	25.8	4.0	0.001	18.1	33.8
	RE	17.0	1.4	0.000	14.3	19.8
	OLS	15.7	1.5	0.000	12.7	18.8