Japanese Journal of Statistics and Data Science (2020) 3:565–581
https://doi.org/10.1007/s42081-020-00084-x

SURVEY ARTICLE

Theory and Practice of Surveys

Empirical likelihood and estimating equations for survey data analysis

Changbao Wu[1] · Mary E. Thompson[1]

Received: 29 February 2020 / Accepted: 3 July 2020 / Published online: 15 July 2020

Abstract
This paper provides an overview of empirical likelihood methods for analysis of survey data when the finite population parameters are defined through a set of census estimating equations. The general inferential framework involving both the superpopulation and the finite population parameters is described, and inferential procedures for point estimation, hypothesis testing, variable selection, and Bayesian analysis, along with the main computational procedures, are discussed.

Keywords Bayesian inference · Complex survey data · Estimating functions · Finite population parameters · Hypothesis testing · Regression analysis · Superpopulation models · Variable selection

1 Estimating equations and empirical likelihood

Maximum-likelihood and least-squares estimation methods are two fundamental pillars of the modern statistical sciences. Suppose that $(y_1, \dots, y_n)$ is an independent and identically distributed (iid) sample from a random variable Y with an assumed parametric distribution $f(y;\theta)$. Under certain regularity conditions, the maximum-likelihood estimator $\hat{\theta}$ of θ, which maximizes the likelihood function $L(\theta) = \prod_{i=1}^{n} f(y_i;\theta)$, is the solution to the score equations:

This research is supported by grants from the Natural Sciences and Engineering Research Council of Canada.

✉ Changbao Wu
cbwu@uwaterloo.ca

Mary E. Thompson
methompson@uwaterloo.ca

[1] Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON, Canada

$$\frac{\partial}{\partial\theta}\log L(\theta) = \sum_{i=1}^{n} \frac{\partial}{\partial\theta}\log f(y_i;\theta) = \mathbf{0}. \tag{1}$$

When the response variable y_i is related to a vector of covariates $\mathbf{x}_i$ and the main objective is to explore relations between y and $\mathbf{x}$, a semiparametric regression model can be specified through the first two conditional moments $E_\xi(y_i \mid \mathbf{x}_i) = \mu(\mathbf{x}_i;\theta)$ and $V_\xi(y_i \mid \mathbf{x}_i) = v_i\sigma^2$, where $\mu(\mathbf{x}_i;\theta)$ is the mean function, which can be linear or nonlinear in the vector of parameters θ, and v_i are known constants which might depend on the given $\mathbf{x}_i$. The notations $E_\xi(\cdot)$ and $V_\xi(\cdot)$ refer to expectation and variance under the assumed semiparametric model, ξ. The weighted least-squares estimator $\hat{\theta}$ of θ, which minimizes the weighted sum of squares of residuals $Q(\theta) = \sum_{i=1}^{n}\{y_i - \mu(\mathbf{x}_i;\theta)\}^2/v_i$, is the solution to the normal equations:

$$\frac{\partial}{\partial\theta}Q(\theta) = -2\sum_{i=1}^{n} \mathbf{D}(\mathbf{x}_i;\theta)v_i^{-1}\{y_i - \mu(\mathbf{x}_i;\theta)\} = \mathbf{0}, \tag{2}$$

where $\mathbf{D}(\mathbf{x}_i;\theta) = \partial\mu(\mathbf{x}_i;\theta)/\partial\theta$. For linear regression models where $\mu(\mathbf{x}_i;\theta) = \mathbf{x}_i'\theta$, we have $\mathbf{D}(\mathbf{x}_i;\theta) = \mathbf{x}_i$. For generalized linear models with $\mu_i = \mu(\mathbf{x}_i;\theta) = \mu(\mathbf{x}_i'\theta)$ and $v_i = v(\mu_i)$, where $\mu(\cdot)$ is a link function and $v(\cdot)$ is a variance function, the solution to (2) is called the quasi-maximum-likelihood estimator of θ (McCullagh and Nelder 1983).

The score Eq. (1) and the normal equations (2) can be unified through a common form:

$$\mathbf{G}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{g}(y_i, \mathbf{x}_i;\theta) = \mathbf{0}, \tag{3}$$

where the estimating functions $\mathbf{g}(y, \mathbf{x};\theta)$ are unbiased, i.e., $E_\xi\{\mathbf{g}(y, \mathbf{x};\theta_0)\} = \mathbf{0}$ under the assumed model, ξ, where θ_0 denotes the true value of θ. The factor $1/n$ in (3) is redundant, but is included, so that the asymptotic order of $\mathbf{G}_n(\theta_0)$ will be $O_p(n^{-1/2})$. Godambe (1960) was the first to study the optimality properties of the score functions given in (1). Some early results on theoretical and applied aspects of estimating functions were collected in the book edited by Godambe (1991).

The general theory of estimating equations has much broader scope than the maximum-likelihood and the least-squares methods. Let $\mathbf{g}(y, \mathbf{x};\theta)$ be a vector of $r \times 1$ estimating functions; let θ be a $k \times 1$ vector of unknown parameters; let $\mathbf{G}(\theta) = E_\xi\{\mathbf{g}(y, \mathbf{x};\theta)\}$ under the assumed model, ξ. The true value θ_0 for the vector parameter satisfies $\mathbf{G}(\theta_0) = \mathbf{0}$. The over-identified scenarios with $r > k$ are often of interest and will be discussed in detail in the paper. For just-identified cases where $r = k$, the so-called m-estimator $\hat{\theta}$ of θ_0 based on the random sample $\{(y_i, \mathbf{x}_i), i = 1, \dots, n\}$ is the solution to $\mathbf{G}_n(\theta) = \mathbf{0}$ as specified by the estimating equations (3). Theoretical properties of the m-estimators with independent samples can be found in Newey and McFadden (1994), van der Vaart (2000), and Tsiatis (2006). Under-identified scenarios with $r < k$ are not of interest for this paper.

Empirical likelihood is one of the major statistical advances of the past 30 years. It was first proposed by Owen (1988) for iid samples. While the development of empirical likelihood has been the collective effort of many contributors, as evidenced in the book by Owen (2001), there were two major milestones that established the approach as a general inference tool. The first major milestone was the result proved by Owen (1988), analogous to Wilks' theorem for parametric models, showing that the nonparametric empirical likelihood ratio statistic has a χ^2 limiting distribution. Let $\mathbf{p} = (p_1, \dots, p_n)$ be the discrete probability measure over the iid sample $(y_1, \dots, y_n)$ from a random variable Y with mean $\mu_0 = E_\xi(Y)$. The distribution of Y based on the sample data is represented by $F_n(t) = \sum_{i=1}^n p_i I(y_i \le t)$, $t \in (-\infty, \infty)$, which is the empirical likelihood estimator of $F(t) = P(Y \le t)$. The maximum value of the empirical likelihood function $L(\mathbf{p}) = \prod_{i=1}^n p_i$ under the normalization constraint:

$$\sum_{i=1}^n p_i = 1 \quad (p_i \ge 0) \tag{4}$$

is achieved at $\hat{p}_i = n^{-1}$, $i = 1, \dots, n$. The maximum empirical likelihood estimator of $F(t)$ reduces to $F_n(t) = n^{-1} \sum_{i=1}^n I(y_i \le t)$, the customary empirical distribution of Y. Let $\hat{p}(\mu)$ be the maximizer of $L(\mathbf{p})$ under the normalization constraint (4) and the constraint induced by the parameter of interest:

$$\sum_{i=1}^n p_i y_i = \mu \tag{5}$$

for a given μ. Owen (1988) showed that, under mild moment conditions on Y, the empirical likelihood ratio statistic $r(\mu) = -2\{\log L(\hat{\mathbf{p}}(\mu)) - \log L(\hat{\mathbf{p}})\}$ converges in distribution to a χ^2 random variable with one degree of freedom when $\mu = \mu_0$.

The second major milestone is the paper by Qin and Lawless (1994) on combining empirical likelihood with general estimating equation theory for parameters defined through unbiased estimating functions. Suppose that the $k \times 1$ vector parameters θ satisfy $E_\xi\{\mathbf{g}(y, \mathbf{x}; \theta)\} = \mathbf{0}$ when $\theta = \theta_0$. The empirical likelihood function for θ is computed as $L(\hat{\mathbf{p}}(\theta))$, where $\hat{\mathbf{p}}(\theta) = (\hat{p}_1(\theta), \dots, \hat{p}_n(\theta))$ maximizes $L(\mathbf{p})$ subject to the normalization constraint (4) and the parameter constraint given by:

$$\sum_{i=1}^n p_i \, \mathbf{g}(y_i, \mathbf{x}_i; \theta) = \mathbf{0} \tag{6}$$

for the given θ. The maximum empirical likelihood estimator $\hat{\theta}$ of θ is obtained as the maximum point of $L(\hat{\mathbf{p}}(\theta))$. There are several impactful consequences from combining estimating equations with empirical likelihood. First, it provides a general approach for dealing with different inferential problems through estimating functions. Second, the $r \times 1$ estimating functions $\mathbf{g}(y, \mathbf{x}; \theta)$ can be over-identified (i.e., $r > k$), which becomes convenient for incorporating auxiliary information and known moment conditions through additional estimating equations. Third, it allows inferences on key parameters of interest while treating others as nuisance parameters. And finally, it opens the door for exploring other advanced inferential procedures such as variable selection and Bayesian analysis through empirical likelihood.

Historically, the same concept of empirical likelihood was first discussed in survey sampling under the name "scale-load approach" by Hartley and Rao (1968, (1969). They focused on point estimation and showed that a constrained maximization problem with the known population mean of the auxiliary variable used in a calibration equation leads to the maximum scale-load estimator which is asymptotically equivalent to the regression estimator. This result was later "re-discovered" by Chen and Qin (1993) using the empirical likelihood formulation of Owen (1988).

2 Design-based inference with survey data

A survey population consists of a finite number N of units. Values of the variables of interest are attached to units and it is assumed that the values are fixed for each unit and can be measured without error. Let $\mathbf{S}$ be the set of n units in the survey sample selected by a probability sampling method. Let $\{(y_i, \mathbf{x}_i), i \in \mathbf{S}\}$ be the survey dataset. We assume that the first-order and the second-order inclusion probabilities π_i and π_{ij} are available, and the survey design leads to a fixed sample size n. Let $d_i = 1/\pi_i$ be the basic survey design weights, $i \in \mathbf{S}$.

Traditional design-based estimation with survey data focuses on descriptive finite population parameters such as the population mean $\mu_y = N^{-1}\sum_{i=1}^N y_i$, the finite population distribution function $F_{\scriptscriptstyle N}(t) = N^{-1}\sum_{i=1}^N I(y_i \le t)$ where $I(\cdot)$ is the indicator function, and the 100αth finite population quantile $t_\alpha = F_{\scriptscriptstyle N}^{-1}(\alpha) = \inf\{t \mid F_{\scriptscriptstyle N}(t) \ge \alpha\}$ with $\alpha \in (0,1)$. The finite population and the finite population parameters are viewed as fixed, and randomization is induced by the probability sampling design for selecting the survey sample. The survey weighted estimator of μ_y is given by $\hat{\mu}_y = \sum_{i\in\mathbf{S}} d_i y_i / \sum_{i\in\mathbf{S}} d_i$, and the estimator of $F_{\scriptscriptstyle N}(t)$ for a given t has the same form of $\hat{\mu}_y$ but with y_i replaced by $I(y_i \le t)$.

Most descriptive finite population parameters can be defined through a (single) census estimating equation in the general form of:

$$\mathbf{G}_{\scriptscriptstyle N}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{g}(y_i, \mathbf{x}_i;\theta) = \mathbf{0}\,. \tag{7}$$

The factor N^{-1} used in (7) as well as (8) below is for the convenience of asymptotic development and is not required for computations. The finite population mean $\theta_{\scriptscriptstyle N} = \mu_y$ corresponds to $\mathbf{g}(y_i, \mathbf{x}_i;\theta) = y_i - \theta$. The 100$\alpha$th finite population quantile $\theta_{\scriptscriptstyle N} = t_\alpha$ is defined through $\mathbf{g}(y_i, \mathbf{x}_i;\theta) = I(y_i \le \theta) - \alpha$. For the population quantiles, the estimating function is not continuous in θ and the equation $\mathbf{G}_{\scriptscriptstyle N}(\theta) = 0$ may not hold exactly for any θ. An alternative solution can be defined through $\theta_{\scriptscriptstyle N} = \inf\{\theta \mid \mathbf{G}_{\scriptscriptstyle N}(\theta) \ge 0\}$, which satisfies $\mathbf{G}_{\scriptscriptstyle N}(\theta_{\scriptscriptstyle N}) = O(N^{-1})$. This modification to (7) does not change the asymptotic results on $\theta_{\scriptscriptstyle N}$. The design-based estimator $\hat{\theta}$ of $\theta_{\scriptscriptstyle N}$ can be obtained as the solution to the survey weighted estimating equation:

$$\mathbf{G}_n(\theta) = \frac{1}{N}\sum_{i\in\mathbf{S}} d_i \mathbf{g}(y_i, \mathbf{x}_i;\theta) = \mathbf{0}\,. \tag{8}$$

Design-based estimation of finite population parameters can be carried out under the unified framework of estimating equations as specified by (7) and (8).

Large-scale complex survey data are often used for inferences on model parameters. This is the so-called analytic use of survey data. The survey variables $(y, \mathbf{x})$ are assumed to follow a model, called the *superpopulation model*, denoted as ξ. The model parameters θ may be defined through a set of unbiased estimating functions, i.e., $\mathbf{G}(\theta) = E_\xi\{\mathbf{g}(y, \mathbf{x}; \theta)\} = \mathbf{0}$. When a conditional model of y given $\mathbf{x}$ is used, the model parameters can be specified through $E_\xi\{\mathbf{g}(y, \mathbf{x}; \theta) \mid \mathbf{x}\} = \mathbf{0}$. One of the statistical questions is how to make inferences on the model parameters θ using a probability survey sample $\mathbf{S}$ selected from a particular finite population. Godambe and Thompson (1986, 2009) proposed to focus on finite population parameters θ_N defined through census estimating equations using design-based methods. If the superpopulation model holds for the survey population and the population size N is large, inferences on θ_N are essentially the same as for the model parameters θ. If the finite population does not follow the model ξ, the finite population parameters θ_N are well defined and may still be of interest for the survey population. Design-based inferences remain valid for the latter cases. We consider two practically important scenarios of the analytic use of survey data.

Linear regression analysis. Suppose that the study variable y and a set of covariates $\mathbf{x}$ are measured for all units in the survey sample, $\mathbf{S}$. For notational simplicity without loss of generality, we assume that the vector $\mathbf{x}$ contains 1 as its first component. The linear regression model is assumed to hold for the finite population, i.e., $y_i = \mathbf{x}_i'\beta + \varepsilon_i, i = 1, \ldots, N$, where the ε_i's are iid with $E_\xi(\varepsilon_i) = 0$ and $V_\xi(\varepsilon_i) = \sigma^2$. The β and σ^2 are the superpopulation parameters. The estimating functions for β under the least-squares estimation framework are given by $\mathbf{g}(y, \mathbf{x}; \beta) = \mathbf{x}(y - \mathbf{x}'\beta)$. The finite population regression coefficients β_N are the solution to $\sum_{i=1}^N \mathbf{x}_i(y_i - \mathbf{x}_i'\beta) = \mathbf{0}$, which leads to the closed form expression $\beta_N = \big(\sum_{i=1}^N \mathbf{x}_i\mathbf{x}_i'\big)^{-1}\sum_{i=1}^N \mathbf{x}_i y_i$. This is the least square estimator of the model parameters β if we treat the finite population as an iid sample of size N from the linear regression model. The survey weighted estimator $\hat{\beta}$ is the solution to $\sum_{i\in\mathbf{S}} d_i\mathbf{x}_i(y_i - \mathbf{x}_i'\beta) = \mathbf{0}$, and is given by $\hat{\beta} = \big(\sum_{i\in\mathbf{S}} d_i\mathbf{x}_i\mathbf{x}_i'\big)^{-1}\sum_{i\in\mathbf{S}} d_i\mathbf{x}_i y_i$.

The linear regression model ξ may not hold for the finite population from which the survey sample is selected. This can happen, for instance, if crucial covariates are not measured by the survey sample or if the model contains certain high order or interaction terms. However, the finite population regression coefficients β_N are still meaningful parameters for the survey population and the design-based estimator $\hat{\beta}$ remains consistent for β_N.

Logistic regression analysis. Suppose that the study variable y is binary and $p_i = P(y_i = 1 \mid \mathbf{x}_i)$ depends on $\mathbf{x}_i$ through the logit link function, i.e., $p_i = p(\mathbf{x}_i'\beta) = 1 - \big\{1 + \exp(\mathbf{x}_i'\beta)\big\}^{-1}$. The estimating functions for the model parameters β under the quasi-maximum likelihood framework of (2) with $v_i = p_i(1 - p_i)$ are given by $\mathbf{g}(y, \mathbf{x}; \beta) = \mathbf{x}\{y - p(\mathbf{x}'\beta)\}$. The finite population regression coefficients β_N under the assumed logistic regression model are the solution to $\sum_{i=1}^N \mathbf{x}_i\{y_i - p(\mathbf{x}_i'\beta)\} = \mathbf{0}$, which does not have a closed form expression. The design-based estimator $\hat{\beta}$ of β_N is the solution to $\sum_{i\in\mathbf{S}} d_i\mathbf{x}_i\{y_i - p(\mathbf{x}_i'\beta)\} = \mathbf{0}$. Finding the solution requires an iterative computational procedure.

When the estimating functions $\mathbf{g}(y,\mathbf{x};\theta)$ are differentiable in θ and the estimating equation system is just-identified (i.e., $r=k$), the design-based estimator $\hat{\theta}$ obtained by solving (8) is design-consistent for θ_N with the design-based variance–covariance matrix given by the sandwich form (Binder 1983):

$$V_p\big(\hat{\theta}\big) = \{\mathbf{H}_N(\theta_N)\}^{-1} V_p\big\{\mathbf{G}_n\big(\theta_N\big)\big\}\{\mathbf{H}'_N(\theta_N)\}^{-1}, \tag{9}$$

where

$$\mathbf{H}_N(\theta) = \frac{\partial}{\partial\theta}\mathbf{G}_N(\theta) = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial\theta}\mathbf{g}(y_i,\mathbf{x}_i;\theta), \tag{10}$$

and $V_p(\cdot)$ denotes variance under the probability sampling design. The design-based variance estimator is computed as:

$$v_p\big(\hat{\theta}\big) = \{\mathbf{H}_n(\hat{\theta})\}^{-1} v_p\big\{\mathbf{G}_n\big(\hat{\theta}\big)\big\}\{\mathbf{H}'_n(\hat{\theta})\}^{-1},$$

where

$$\mathbf{H}_n(\theta) = \frac{\partial}{\partial\theta}\mathbf{G}_n(\theta) = \frac{1}{N}\sum_{i\in\mathbf{S}} d_i\left\{\frac{\partial}{\partial\theta}\mathbf{g}(y_i,\mathbf{x}_i;\theta)\right\}, \tag{11}$$

and $v_p\big\{\mathbf{G}_n\big(\hat{\theta}\big)\big\}$ is the design-based estimator of the variance-covariance matrix for the Horvitz–Thompson estimator $\mathbf{G}_n\big(\theta\big)$ evaluated at $\theta=\hat{\theta}$.

3 General inferential procedures

In this section, we discuss two empirical likelihood-based inferential problems for parameters θ_N defined through the census estimating equations (7). We consider the general setting where $r \ge k$ and the estimating functions $\mathbf{g}(y,\mathbf{x};\theta)$ can be smooth or non-differentiable. The asymptotic framework assumes that there is a sequence of finite populations and a sequence of probability survey samples, indexed by ν. Both the population size N_ν and the sample size n_ν go to infinity as $\nu \to \infty$. All limiting processes are understood as $\nu \to \infty$; see Fuller (2009) for further details. The index ν will be dropped for notational simplicity and the limiting processes are denoted exchangeably as $N \to \infty$ or $n \to \infty$.

3.1 Empirical likelihood-based inferences with survey data

Standard empirical likelihood methods for independent sample data with parameters defined through estimating equations consist of three main components: (i) the empirical likelihood function $L(\mathbf{p}) = \prod_{i=1}^{n} p_i$; (ii) the normalization constraint (4); and (iii) the parameter constraints (6). When the methods are applied directly to survey data, the resulting estimator $\hat{\theta}$ is not design-consistent unless the sample is selected by simple random sampling. There are two possible modifications to make the methods applicable to survey data analysis. One is to modify the empirical likelihood function $L(\mathbf{p})$ to take into account the survey design features, and the other is to use a survey weighted version for the parameter constraints.

Pseudo empirical likelihood methods. Chen and Sitter (1999) proposed to replace the empirical log-likelihood function $\ell(\mathbf{p}) = \sum_{i=1}^{n} \log(p_i)$ by the pseudo empirical log-likelihood function $\ell_{PEL}(\mathbf{p}) = \sum_{i\in\mathbf{S}} d_i \log(p_i)$ while keeping the normalization constraint (4) and the parameter constraints (6) unchanged. The method leads to design-consistent point estimators. Pseudo empirical likelihood ratio confidence intervals were discussed by Wu and Rao (2006) for a scalar parameter. Generalizations to vector parameters defined through estimating equations were given in Zhao and Wu (2019).

Sample empirical likelihood methods. The sample empirical likelihood was first briefly mentioned by Chen and Kim (2014) as an alternative approach to the population empirical likelihood methods discussed in their paper. The methods were formally studied by Zhao et al. (2019) and Zhao and Wu (2019). The sample empirical likelihood uses the same form $L(\mathbf{p})$ as for iid data and the standard normalization constraint (4), but replaces the parameter constraints (6) by a survey weighted version. These methods also lead to design-consistent point estimators.

Our discussions for the rest of the paper are formulated under the sample empirical likelihood. The empirical likelihood methods discussed by Berger and De La Riva Torres (2016) and Oguz-Alper and Berger (2016) are also closely related to the sample empirical likelihood methods.

3.2 Point estimation

We first consider point estimation for finite population parameters θ_N defined through the census estimating equations (7). The sample empirical log-likelihood function is given by $\ell_{SEL}(\mathbf{p}) = \sum_{i\in\mathbf{S}} \log(p_i)$. The sample empirical likelihood function of θ is defined as:

$$\ell_{SEL}(\theta) = \ell_{SEL}\{\hat{\mathbf{p}}(\theta)\} = \sum_{i\in\mathbf{S}} \log\{\hat{p}_i(\theta)\}\,, \tag{12}$$

where $\hat{\mathbf{p}}(\theta) = (\hat{p}_1(\theta), \ldots, \hat{p}_n(\theta))$ maximizes $\ell_{SEL}(\mathbf{p}) = \sum_{i\in\mathbf{S}} \log(p_i)$ subject to the normalization constraint $\sum_{i\in\mathbf{S}} p_i = 1$ and the survey weighted parameter constraints:

$$\sum_{i\in\mathbf{S}} p_i\{d_i\mathbf{g}(y_i, \mathbf{x}_i; \theta)\} = \mathbf{0} \tag{13}$$

for the given θ. The maximum sample empirical likelihood estimator $\hat{\theta}$ of θ_N is the maximum point of $\ell_{SEL}(\theta)$, i.e., $\hat{\theta} = \arg\max_{\theta\in\Theta} \ell_{SEL}(\theta)$, where Θ is the parameter space.

The design-based validity of the maximum sample empirical likelihood estimator $\hat{\theta}$ can be informally justified by two special cases. When the estimating equations system (7) is just-identified (i.e., $r = k$), the global maximum of $\ell_{SEL}(\mathbf{p})$ is achieved at $\hat{p}_i = n^{-1}$ for all $i \in \mathbf{S}$, and the maximum sample empirical likelihood estimator $\hat{\theta}$ is the solution to the survey weighted estimating equations (8), which is design-consistent under suitable regularity conditions. A practically important over-identified estimating equation system is the use of known auxiliary population information for survey data analysis. Let $\mathbf{g}(y, \mathbf{x}, \mathbf{z}; \theta) = (\mathbf{g}_1'(y, \mathbf{x}; \theta), \mathbf{g}_2'(\mathbf{z}))'$, where $\mathbf{g}_1(y, \mathbf{x}; \theta)$ are the $k \times 1$ estimating functions for defining the $k \times 1$ parameters $\theta_{_N}$, and $\mathbf{g}_2(\mathbf{z})$ are $(r-k) \times 1$ estimating functions which do not involve the parameters θ and satisfy the moment condition $N^{-1} \sum_{i=1}^{N} \mathbf{g}_2(\mathbf{z}_i) = \mathbf{0}$. For instance, we may have $\mathbf{g}_2(\mathbf{z}_i) = \mathbf{z}_i - \mu_{\mathbf{z}}$ where the finite population means $\mu_{\mathbf{z}}$ for the $\mathbf{z}$ variables are known and can be used in benchmark constraints. The parameter constraints under the current setting are given by:

$$\sum_{i \in \mathbf{S}} p_i \{d_i \mathbf{g}(y_i, \mathbf{x}_i, \mathbf{z}_i; \theta)\} = \mathbf{0}\,,$$

which is an over-identified system. It can be shown that the maximum sample empirical likelihood estimator $\hat{\theta}$ solves the first part of the just-identified equations system:

$$\sum_{i \in \mathbf{S}} \hat{p}_i \{d_i \mathbf{g}_1(y_i, \mathbf{x}_i; \theta)\} = \mathbf{0}\,, \tag{14}$$

where $\hat{\mathbf{p}} = (\hat{p}_1, \ldots, \hat{p}_n)$ is the maximizer of $\ell_{_{SEL}}(\mathbf{p})$ under the normalization constraint (4) and the benchmark constraints (second part of the equations system):

$$\sum_{i \in \mathbf{S}} p_i \{d_i \mathbf{g}_2(\mathbf{z}_i)\} = \mathbf{0}\,.$$

The combined components $\hat{p}_i d_i$ can be viewed as the calibration weights, and the solution $\hat{\theta}$ to the estimating equations (14) is design-consistent for $\theta_{_N}$ defined through $N^{-1} \sum_{i=1}^{N} \mathbf{g}_1(y_i, \mathbf{x}_i; \theta) = \mathbf{0}$.

An over-identified estimating equation system does not always have a partition $(\mathbf{g}_1, \mathbf{g}_2)$ with the calibration equations described above. For instance, if the parameter θ is the mean of a Poisson random variable y, then the single θ satisfies two moment conditions: $E_\xi(y - \theta) = 0$ and $E_\xi\{(y-\theta)^2 - \theta\} = 0$. In another example, if θ is the mean of the variable y with a known variance σ_0^2, then the parameter θ also satisfies two moment conditions: $E_\xi(y - \theta) = 0$ and $E_\xi\{(y-\theta)^2 - \sigma_0^2\} = 0$. General estimation results which cover over-identified estimating equations system are both theoretically and practically important.

Under suitable regularity conditions on the estimating functions $\mathbf{g}(y, \mathbf{x}; \theta)$, the probability sampling design, and the finite population as described in Zhao et al. (2019), the maximum sample empirical likelihood estimator $\hat{\theta}$ is design-consistent with design-based variance–covariance matrix given by:

$$\mathbf{V} = \left(\mathbf{H}'\mathbf{W}^{-1}\mathbf{H}\right)^{-1} \mathbf{H}'\mathbf{W}^{-1}\boldsymbol{\Sigma}\mathbf{W}^{-1}\mathbf{H}\left(\mathbf{H}'\mathbf{W}^{-1}\mathbf{H}\right)^{-1}, \tag{15}$$

where $\mathbf{H} = \mathbf{H}_{_N}(\theta_{_N})$ and $\mathbf{H}_{_N}(\theta)$ is defined in (10), $\mathbf{W} = nN^{-2} \sum_{i=1}^{N} d_i \mathbf{g}_i \mathbf{g}_i'$ with $\mathbf{g}_i = \mathbf{g}(y_i, \mathbf{x}_i; \theta_{_N})$, and $\boldsymbol{\Sigma} = V_p\left\{\mathbf{G}_n\left(\theta_{_N}\right)\right\}$ as previously appeared in (9). It should be noted that $\mathbf{H}$ is $r \times k$, $\mathbf{W}$ is $r \times r$, and $\boldsymbol{\Sigma}$ is $r \times r$, resulting in a $k \times k$ matrix for $\mathbf{V}$.

If the estimating equation system is just-identified (i.e., $r = k$), the variance–covariance matrix given in (15) reduces to $\mathbf{V} = \mathbf{H}^{-1}\boldsymbol{\Sigma}(\mathbf{H}')^{-1}$, which is the same as $V_p(\hat{\theta})$ given in (9). In general, variance estimation requires plug-in estimators for the three components $\mathbf{H}$, $\mathbf{W}$ and $\boldsymbol{\Sigma}$, which are, respectively, given by $\hat{\mathbf{H}} = \mathbf{H}_n(\hat{\theta})$ as defined in (11), $\hat{\mathbf{W}} = nN^{-2}\sum_{i\in\mathbf{S}} d_i^2 \hat{\mathbf{g}}_i \hat{\mathbf{g}}_i'$ with $\hat{\mathbf{g}}_i = \mathbf{g}(y_i, \mathbf{x}_i; \hat{\theta})$, and $\hat{\boldsymbol{\Sigma}} = v_p\left\{\mathbf{G}_n\left(\hat{\theta}\right)\right\}$. The definitions of $\mathbf{H}$ and $\hat{\mathbf{H}}$ through (10) and (11) cannot be used when the estimating functions $\mathbf{g}_i = \mathbf{g}(y_i, \mathbf{x}_i; \theta)$ are non-differentiable in θ. The asymptotic result under those cases involves $\mathbf{H}(\theta) = \partial\mathbf{G}(\theta)/\partial\theta$, where $\mathbf{G}(\theta) = \lim_{N\to\infty}\mathbf{G}_N(\theta)$. Zhao and Wu (2019) contains details on how to estimate $\mathbf{H}$ for non-smooth estimating functions and additional discussions on estimating the design-based variance–covariance matrix $\boldsymbol{\Sigma}$ under commonly used survey designs.

3.3 Hypothesis tests

Hypothesis tests are a common inferential problem for building statistical models or answering specific scientific questions. With complex survey data, the problems can be formulated for finite population parameters defined through census estimating equations under the design-based framework. When the assumed superpopulation model holds for the survey population, the inferential results can be extended to the superpopulation model parameters as discussed in Sect. 2.

The general results on sample empirical likelihood ratio tests and the required regularity conditions are discussed in Zhao et al. (2019) and Zhao and Wu (2019). The sample empirical likelihood ratio statistic for testing $H_0 : \theta_N = \theta_{N0}$ versus $H_1 : \theta_N \neq \theta_{N0}$ for a pre-specified θ_{N0} is computed as:

$$r_{SEL}(\theta_{N0}) = -2\left\{\ell_{SEL}(\theta_{N0}) - \ell_{SEL}(\hat{\theta})\right\},$$

where $\ell_{SEL}(\theta)$ is defined in (12) and $\hat{\theta}$ is the maximum sample empirical likelihood estimator of θ_N. It can be shown that:

$$r_{SEL}(\theta_{N0}) = \mathbf{Q}'\boldsymbol{\Delta}\mathbf{Q} + o_p(1),$$

where $\mathbf{Q} \sim \mathbf{N}(\mathbf{0}, \mathbf{I}_r)$, the standard multivariate normal distribution, and:

$$\boldsymbol{\Delta} = n\boldsymbol{\Sigma}^{1/2}\mathbf{W}^{-1}\mathbf{H}\left(\mathbf{H}'\mathbf{W}^{-1}\mathbf{H}\right)^{-1}\mathbf{H}'\mathbf{W}^{-1}\boldsymbol{\Sigma}^{1/2}.$$

The sampling distribution of $r_{SEL}(\theta_{N0})$ is asymptotically equivalent to the distribution of a quadratic form, which can be re-expressed as:

$$\mathbf{Q}'\boldsymbol{\Delta}\mathbf{Q} = \sum_{j=1}^{k}\delta_j\chi_j^2,$$

where δ_j, $j = 1, \ldots, k$ are the non-zero eigenvalues of $\boldsymbol{\Delta}$, and χ_j^2, $j = 1, \ldots, k$ are independent χ^2 random variables with one degree of freedom. For just-identified cases with $r = k$, the matrix $\boldsymbol{\Delta}$ reduces to $\boldsymbol{\Delta} = n\boldsymbol{\Sigma}^{1/2}\mathbf{W}^{-1}\boldsymbol{\Sigma}^{1/2}$. It can further be shown that, under single-stage PPS sampling with a negligible sampling fraction, we have $\boldsymbol{\Sigma} = n^{-1}\mathbf{W} + o(n^{-1})$, and consequently, the sample empirical likelihood ratio statistic $r_{SEL}(\theta_{N0})$ converges in distribution to a standard χ^2 random variable with k degrees of freedom.

The sample empirical likelihood ratio statistic $r_{SEL}(\theta_N \mid H_0)$ for testing a general hypothesis $H_0 : \mathbf{K}(\theta_N) = \mathbf{0}$ versus $H_1 : \mathbf{K}(\theta_N) \neq \mathbf{0}$, where $\mathbf{K}(\theta_N) = \mathbf{0}$ imposes k_1 $(\leq k)$ linear or nonlinear constraints on the $k \times 1$ parameters θ_N, is computed as follows. Let $\hat{\theta}$ be the (unrestricted) maximum sample empirical likelihood estimator of θ_N over the parameter space Θ; let $\hat{\theta}^* = \arg\max_{\theta \in \Theta^*} \ell_{SEL}(\theta)$ be the restricted maximum sample empirical likelihood estimator of θ_N under the restricted parameter space $\Theta^* = \{\theta \mid \theta \in \Theta \text{ and } \mathbf{K}(\theta) = \mathbf{0}\}$. We have:

$$r_{SEL}(\theta_N \mid H_0) = -2\Big\{\ell_{SEL}(\hat{\theta}^*) - \ell_{SEL}(\hat{\theta})\Big\}.$$

It can be shown that $r_{SEL}(\theta_N \mid H_0) = \mathbf{Q}'\boldsymbol{\Delta}^*\mathbf{Q} + o_p(1)$, where $\mathbf{Q} \sim \mathbf{N}(\mathbf{0}, \mathbf{I}_r)$ and:

$$\boldsymbol{\Delta}^* = n\boldsymbol{\Sigma}^{1/2}\mathbf{W}^{-1}\mathbf{H}\boldsymbol{\Gamma}\boldsymbol{\Phi}'\left(\boldsymbol{\Phi}\boldsymbol{\Gamma}\boldsymbol{\Phi}'\right)^{-1}\boldsymbol{\Phi}\boldsymbol{\Gamma}\mathbf{H}'\mathbf{W}^{-1}\boldsymbol{\Sigma}^{1/2},$$

where $\boldsymbol{\Phi} = \{\partial\mathbf{K}(\theta)/\partial\theta\}|_{\theta=\theta_N}$ and $\boldsymbol{\Gamma} = \mathbf{H}'\mathbf{W}^{-1}\mathbf{H}$, with $\mathbf{H}$, $\mathbf{W}$ and $\boldsymbol{\Sigma}$ defined the same as before. If $r = k$ and the survey design is single-stage PPS sampling with a small sampling fraction, the sample empirical likelihood ratio statistic $r_{SEL}(\theta_N \mid H_0)$ follows asymptotically a standard χ^2 distribution with k_1 degrees of freedom.

Linear hypotheses are most commonly encountered in practice, where $\mathbf{K}(\theta) = \mathbf{0}$ has the form $\mathbf{A}\theta = \mathbf{b}$, with $\mathbf{A}$ being a $k_1 \times k$ matrix and $\mathbf{b}$ being a $k_1 \times 1$ vector, both pre-specified. In this case, we have $\boldsymbol{\Phi} = \partial\mathbf{K}(\theta)/\partial\theta = \mathbf{A}$. The hypothesis $H_0 : \theta_N = \theta_{N0}$ is equivalent to letting $\mathbf{A} = \mathbf{I}_k$ and $\mathbf{b} = \theta_{N0}$.

Implementations of the sample empirical likelihood ratio tests generally require the estimation of the matrix $\boldsymbol{\Delta}$ or $\boldsymbol{\Delta}^*$, which amounts to estimating $\mathbf{H}$, $\mathbf{W}$ and $\boldsymbol{\Sigma}$. The sampling distribution of the test statistic $r_{SEL}(\theta_{N0})$ or $r_{SEL}(\theta_N \mid H_0)$ can be obtained through a simulation-based approach to the distribution of the quadratic form $\mathbf{Q}'\boldsymbol{\Delta}\mathbf{Q}$ or $\mathbf{Q}'\boldsymbol{\Delta}^*\mathbf{Q}$, or equivalently, the linear combination of independent χ^2 random variables using the estimated eigenvalues of $\boldsymbol{\Delta}$ or $\boldsymbol{\Delta}^*$. Some analytic approximation methods for the distribution of a weighted sum of χ^2 random variables such as those described in Rao and Scott (1981) and Rao and Scott (1984) and Bodenham and Adams (2016) may also be considered.

4 Design-based variable selection

Complex survey data often contain information on a large number of variables, especially for health and social science-related surveys where many factors are deemed potentially important for scientific investigations. For instance, surveys of the International Tobacco Control (ITC) Policy Evaluation Project (Thompson et al. 2006) collect data on many variables related to demographic, psychosocial, behavioral, and health aspects of the units as well as measures of knowledge and attitude towards smoking. Variable selection is an important problem at the initial stage of model building to identify relevant factors for a particular response variable such as addiction or quitting behaviors.

Design-based variable selection using survey data focuses on the finite population regression coefficients for linear regression models, logistic regression models, or other generalized linear models as discussed in Sect. 2. Under standard settings with independent sample data, the basic aim of variable selection is to identify covariates in a regression model for which the coefficients are zero. The finite population regression coefficients θ_N defined as the solution to the census estimating equations, however, are usually not exactly equal to zero even if the corresponding superpopulation parameters are zero. The components are typically of the order $O(N^{-1/2})$ if the model parameters are zero and the model holds for the finite population. For design-based variable selection, we need to treat population regression coefficients as practically zero if their theoretical values are of the order $O(N^{-1/2})$.

The most widely known variable selection method that is a product of an estimation technique is the *least absolute shrinkage and selection operator* (LASSO) by Tibshirani (1996). Variable selection through penalized empirical likelihood with independent data has been studied by Tang and Leng (2010) and Leng and Tang (2012). The general procedures require that the un-penalized method provides consistent point estimators of the regression coefficients, and the penalized method forces estimators with small values to be zero. The sample empirical likelihood fits into this framework very naturally for design-based variable selection with the population regression coefficients defined through census estimating equations.

Let $p_\tau(\cdot)$ be a pre-specified penalty function with regularization parameter τ. Let $\mathbf{g}(y, \mathbf{x};\theta)$ be the estimating functions for defining θ_N. The penalized sample empirical likelihood function (omitting the constant term $-n\log(n)$) is defined as:

$$\ell_{PSEL}(\theta) = -\sum_{i\in\mathbf{S}} \log\left[1 + \lambda'\{d_i\mathbf{g}(y_i, \mathbf{x}_i;\theta)\}\right] - n\sum_{j=1}^{k} p_\tau(|\theta_j|),$$

where θ_j is the jth component of θ and the Lagrange multiplier λ is the solution to (17) as described in Sect. 6. The *smoothly clipped absolute deviation* (SCAD) penalty function proposed by Fan and Li (2001) has been shown to achieve variable selection and unbiased parameter estimation simultaneously under standard settings. Zhao et al. (2019) showed that the SCAD penalty also works well for the penalized sample empirical likelihood method. The SCAD penalty function $p_\tau(t)$ satisfies $p_\tau(0) = 0$ and has its first-order derivative given by:

$$p'_\tau(t) = \tau\left\{ I(t \le \tau) + \frac{(a\tau - t)_+}{(a-1)\tau} I(t > \tau) \right\},$$

where $(b)_+ = b$ if $b \ge 0$ and $(b)_+ = 0$ if $b < 0$. The penalty function contains two regularization parameters: a and τ. The choice $a = 3.7$ works well under the universal thresholding $\tau = \{2\log(k)\}^{1/2}$ when $k \le 100$. More refined data-driven choice of (a, τ) can be determined using criteria such as BIC or generalized cross-validation. See Fan and Li (2001) and Tang and Leng (2010) for further details.

The maximum penalized sample empirical likelihood estimator of θ_N is given by $\hat{\theta}_{PSEL} = \arg\max_{\theta\in\Theta} \ell_{PSEL}(\theta)$. Zhao et al. (2019) showed that the procedure possesses oracle properties for variable selection under the design-based framework in the sense that zero components of θ_N will be correctly identified with probability approaching 1 as n grows large. In addition, the penalized estimator for the non-zero components of θ_N is design-consistent.

5 Bayesian inferences

Bayesian inferences require a likelihood function for the observed sample data. With a chosen prior distribution, inferences on the parameters based on the posterior distribution are conditional on the given sample data. Bayesian inferences for finite population parameters with desirable frequentist properties under the design-based framework, however, are very difficult to achieve, as shown by Godambe (1966, 1968) and Ericson (1969).

The sample empirical likelihood provides a convenient tool for defining a profile likelihood function for finite population parameters through survey weighted estimating equations. With a suitably chosen prior distribution, the likelihood leads to Bayesian posterior inferences which are valid under the design-based framework under certain survey designs. The approach is particularly appealing for parameters involving non-smooth estimating functions, since the computational procedures do not incur any additional difficulties. Upon omitting the constant term $-n\log(n)$, the profile sample empirical log-likelihood function for θ defined in (12) is given by:

$$\ell(\theta) = -\sum_{i\in\mathbf{S}} \log\left[1 + \lambda'\left\{d_i\mathbf{g}(y_i, \mathbf{x}_i;\theta)\right\}\right],$$

where the Lagrange multiplier $\lambda = \lambda(\theta)$ with the given θ is the solution to (17). The maximum sample empirical likelihood estimator $\hat{\theta}$ is the maximum point of $\ell(\theta)$.

5.1 Bayesian inference with a fixed prior

Let $\mathbf{g}_i(\theta) = \mathbf{g}(y_i, \mathbf{x}_i;\theta)$; let $\pi(\theta)$ be a fixed prior distribution which is independent of the sample size n. The posterior distribution of θ for the given sample $\mathbf{S}$ has the form $\pi(\theta \mid \mathbf{S}) \propto \pi(\theta)\exp\{\ell(\theta)\}$ and is given by:

$$\pi(\theta \mid \mathbf{S}) = c(\mathbf{S})\exp\left[\log\left\{\pi(\theta)\right\} - \sum_{i\in\mathbf{S}}\log\left\{1 + \lambda' d_i\mathbf{g}_i(\theta)\right\}\right], \tag{16}$$

where $c(\mathbf{S})$ is a normalizing constant depending on $\{(y_i, \mathbf{x}_i, d_i), i \in \mathbf{S}\}$, such that $\int \pi(\theta \mid \mathbf{S})d\theta = 1$.

It is shown by Zhao et al. (2020) that the posterior density function given in (16) with a fixed prior has the following asymptotic expansion:

$$\pi(\theta \mid \mathbf{S}) \propto \exp\left[-\frac{1}{2}\left(\theta - \hat{\theta}\right)'\mathbf{J}_n\left(\theta - \hat{\theta}\right) + R_n\right],$$

where $\mathbf{J}_n = n\mathbf{H}'\mathbf{W}^{-1}\mathbf{H}$ and $R_n = o_p(1)$, with $\mathbf{H}$ and $\mathbf{W}$ defined the same as before. The posterior distribution of θ is asymptotically equivalent to a multivariate normal distribution with mean $\hat{\theta}$ and variance–covariance matrix $\mathbf{J}_n^{-1}$. The fixed prior distribution $\pi(\theta)$ has no impact on the posterior distribution under large samples.

The asymptotic expansion of the posterior density function shows that the posterior variance of θ matches the design-based variance of the posterior mean under single-stage PPS sampling without replacement with negligible sampling fractions. Consequently, Bayesian inference with any fixed prior has valid design-based frequentist properties under such survey designs.

5.2 Bayesian inference with an *n*-dependent prior

A fixed prior has impact on the analysis when the sample size is small or moderate, but the influence diminishes under large samples. A stronger version of prior distributions is the so-called n-dependent prior, denoted as $\pi_n(\theta)$, for which the variance of the prior distribution shrinks as n gets large. There are practical scenarios where an n-dependent prior might arise naturally. For instance, a previous survey or a pilot survey might be available, which is taken from the same finite population with a common set of variables to those of the current survey. It is possible to obtain a point estimate with an estimated variance from the survey for the parameters of interest, and using the estimates to form a prior distribution. This was used by Rao and Ghangurde (1972) for Bayesian optimization in sampling finite populations.

The n-dependent prior $\pi_n(\theta)$ is assumed to satisfy that (i) the function $\log\{\pi_n(\theta)\}$ is twice continuously differentiable; (ii) the prior density has bounded mode $\mathbf{m}_0 = \arg\max_\theta \pi_n(\theta)$; and (iii) the information matrix satisfies:

$$\mathbf{H}_0 = -\left[\frac{\partial^2}{\partial\theta\partial\theta'}\log\left\{\pi_n(\theta)\right\}\right]\Big|_{\theta=\mathbf{m}_0} = O(n).$$

It is shown by Zhao et al. (2020) that the posterior density $\pi(\theta \mid \mathbf{S})$ given in (16) but with the n-dependent prior $\pi_n(\theta)$ has the following asymptotic expansion:

$$\pi(\theta \mid \mathbf{S}) \propto \exp\left[-\frac{1}{2}\left(\theta - \mathbf{m}_n\right)'\mathbf{K}_n\left(\theta - \mathbf{m}_n\right) + R_n\right],$$

where $\mathbf{K}_n = \mathbf{H}_0 + \mathbf{J}_n^{-1}$, $\mathbf{m}_n = \mathbf{K}_n^{-1}\left(\mathbf{H}_0\mathbf{m}_0 + \mathbf{J}_n^{-1}\hat{\theta}\right)$, $R_n = o_p(1)$, and $\mathbf{J}_n$ is defined in Sect. 5.1. The posterior distribution of θ is asymptotically equivalent to a multivariate normal distribution, of which the mean is a convex combination of the prior mode $\mathbf{m}_0$ and the maximum sample empirical likelihood estimator $\hat{\theta}$, and the variance is inversely related to the sum of the information matrix of the prior and the posterior variance under the noninformative prior.

The asymptotic expansion of the posterior density with an n-dependent prior shows that the impact of the prior distribution is asymptotically negligible if the information matrix of the prior satisfies $\mathbf{H}_0 = o(n)$. This leads to another crucial observation: the condition $\mathbf{m}_0 = \theta_N + O_p(n^{-1/2})$ on the prior distribution is necessary for the validity of design-based frequentist interpretation for Bayesian inference if the variance of the prior distribution is chosen with the order $O(n^{-1})$. If the variance of the prior distribution goes to 0 faster than n^{-1}, the posterior mean will be dominated by the prior mean under large samples. For finite samples, the impact of the n-dependent prior $\pi_n(\theta)$ depends largely on the mode $\mathbf{m}_0$ of the distribution and, to a lesser extent, on the variance of the distribution or the information matrix $\mathbf{H}_0$.

6 Computational notes

The first major computational task is to maximize $\ell_{SEL}(\mathbf{p}) = \sum_{i\in\mathbf{S}} \log(p_i)$ under the constraints (4) and (13) with a given θ. It can be shown using the Lagrange multiplier method that the solution is given by:

$$\hat{p}_i(\theta) = \frac{1}{n\left[1 + \lambda'\left\{d_i\mathbf{g}(y_i, \mathbf{x}_i;\theta)\right\}\right]}$$

for $i \in \mathbf{S}$, where the Lagrange multiplier $\lambda = \lambda(\theta)$, which depends on θ, is the solution to:

$$\mathbf{D}_1(\theta, \lambda) = \frac{1}{n}\sum_{i\in\mathbf{S}} \frac{d_i\mathbf{g}(y_i, \mathbf{x}_i;\theta)}{1 + \lambda'\left\{d_i\mathbf{g}(y_i, \mathbf{x}_i;\theta)\right\}} = \mathbf{0}. \qquad (17)$$

The modified Newton–Raphson procedure proposed by Chen et al. (2002) is designed to solve (17) to obtain λ with a given θ.

The second major computational task is to find the maximum sample empirical likelihood estimator $\hat{\theta} = \arg\max_{\theta\in\Theta} \ell_{SEL}(\theta)$. It can be shown that setting $\partial\ell_{SEL}(\theta)/\partial\theta = \mathbf{0}$ leads to:

$$\mathbf{D}_2(\theta, \lambda) = \left\{\sum_{i\in\mathbf{S}} \hat{p}_i(\theta) d_i \frac{\partial}{\partial\theta}\mathbf{g}(y_i, \mathbf{x}_i;\theta)\right\}' \lambda = \mathbf{0}. \qquad (18)$$

Note that $\mathbf{D}_1(\theta, \lambda)$ and λ are both $r \times 1$, and $\mathbf{D}_2(\theta, \lambda)$ and θ are both $k \times 1$. The estimator $\hat{\theta}$ can be obtained by treating θ and λ as separate parameters and solving (17) and (18) simultaneously.

Variable selection using penalized sample empirical likelihood requires maximization of the penalized sample empirical likelihood $\ell_{PSEL}(\theta)$ with respect to θ. The SCAD penalty function of Fan and Li (2001) allows for a quadratic approximation given by:

$$p_\tau(|\theta_j|) \doteq p_\tau(|\theta_{j0}|) + \frac{1}{2}\left\{p'_\tau(|\theta_{j0}|)/|\theta_{j0}|\right\}(\theta_j^2 - \theta_{j0}^2),$$

when θ_j is close to θ_{j0}, which is an important feature for easy computation. We can replace (18) by $\partial\ell_{PSEL}(\theta)/\partial\theta = \mathbf{0}$ using the quadratic approximation.

Bayesian inferences based on the posterior distribution $\pi(\theta \mid \mathbf{S})$ given in (16) can be carried out through an MCMC procedure. The full posterior distribution of θ_N can be simulated using an acceptance–rejection sampling method. Details can be found in Zhao et al. (2020).

7 Additional remarks

Empirical likelihood and estimating equations are a powerful statistical tool for data analysis. Their applications to survey data analysis require careful adaptations to take account of the survey design features under a suitable framework. Chapters 7 and 8 of Wu and Thompson (2020) contain additional materials on regression analysis, estimating equations, and empirical likelihood with complex survey data.

The pseudo empirical likelihood and the sample empirical likelihood approaches can be applied to survey data with a complex design involving stratification, clustering, and unequal probability selection as characterized by the first- and the second-order inclusion probabilities. The formulation of the sample empirical likelihood through survey weighted estimating equations only involves the first-order inclusion probabilities. This is sufficient for point estimation. Tests of statistical hypotheses require variance estimation, which further requires the second-order inclusion probabilities unless the survey design permits variance approximations without such information. A reviewer raised the interesting question of two-phase sampling designs, where a large first phase sample with information on auxiliary variables is available. Applications of the empirical likelihood methods to two-phase survey data require a careful formulation of constraints similar to those presented in Wu and Luan (2003). They also require detailed derivations of the variance components under two-phase sampling.

Large-scale survey data are often made available to public users who explore different aspects of the data. Public-use survey data files are created to include the basic design weights or the calibration weights as a separate column in addition to all other variables measured by the survey. Variance estimation is typically handled by using additional columns of replication weights supplied by the data file producers. If such weights are available, the inferential procedures described in this paper can readily be applied to public-use survey data files. Zhao et al. (2020) contains further details on empirical likelihood methods with public-use survey data.

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Berger, Y. G., & De La Riva Torres, O. (2016). Empirical likelihood confidence intervals for complex sampling designs. *Journal of the Royal Statistical Society, Series B*, *78*, 319–341.

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

Bodenham, D. A., & Adams, N. M. (2016). A comparison of efficient approximations for a weighted sum of chi-squared random variables. *Statistics and Computing*, *26*, 917–928.

Chen, J., & Qin, J. (1993). Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. *Biometrika*, *80*, 107–116.

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

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

Chen, S., & Kim, J. K. (2014). Population empirical likelihood for nonparametric inference in survey sampling. *Statistica Sinica*, *24*, 335–355.

Ericson, W. A. (1969). Subjective Bayesian models in sampling finite populations, I. *Journal of the Royal Statistical Society, Series B*, *31*, 195–234.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*, 1348–60.

Fuller, W. A. (2009). *Sampling Statistics*. Hoboken, NewJersey: Wiley.

Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *Annals of Mathematical Statistics*, *31*, 1208–1212.

Godambe, V. P. (1966). A new approach to sampling from finite populations. *Journal of the Royal Statistical Society, Series B*, *28*, 310–328.

Godambe, V. P. (1968). Bayesian sufficiency in survey-sampling. *Annals of Mathematical Statistics*, *20*, 363–373.

Godambe, V. P. (1991). *Estimating Functions*. New York: Oxford University Press.

Godambe, V. P., & Thompson, M. E. (1986). Parameters of superpopulation and survey population: Their relationships and estimation. *International Statistical Review*, *54*, 127–138.

Godambe, V.P., & Thompson, M.E. (2009). Estimating functions and survey sampling. In: D. Pfeffermann & C.R. Rao (eds.) *Handbook of Statistics, Volume 29B, Sample Surveys: Inference and Analysis* (pp. 83–101).

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

Hartley, H. O., & Rao, J. N. K. (1969). A new estimation theory for sample surveys, II. In N. L. Johnson & H. Smith (Eds.), *New Developments in Survey Sampling* (pp. 147–169). New York: Wiley.

Leng, C., & Tang, C. Y. (2012). Penalized empirical likelihood and growing dimensional general estimating equations. *Biometrika*, *99*, 703–716.

McCullagh, P., & Nelder, J. (1983). *Generalized Linear Models*. London: Chapman and Hall.

Newey, W. K., & McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of Econometrics*, *4*, 2111–2245.

Oguz-Alper, M., & Berger, Y. G. (2016). Modelling complex survey data with population level information: an empirical likelihood approach. *Biometrika*, *103*, 447–459.

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, *75*, 237–249.

Owen, A. B. (2001). *Empirical Likelihood*. New York: Chapman&Hall/CRC.

Qin, J., & Lawless, J. F. (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics*, *22*, 300–325.

Rao, J. N. K., & Ghangurde, P. D. (1972). Bayesian optimization in sampling finite populations. *Journal of the American Statistical Association*, *67*, 439–443.

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

Rao, J. N. K., & Scott, A. (1984). On chi-squared tests for multi-way tables with cell proportions estimated from survey data. *Annals of Statistics*, *12*, 46–60.

Rao, J. N. K., & Wu, C. (2010). Bayesian pseudo empirical likelihood intervals for complex surveys. *Journal of the Royal Statistical Society, Series B*, *72*, 533–544.

Tang, C. Y., & Leng, C. (2010). Penalized high dimensional empirical likelihood. *Biometrika*, *97*, 905–920.

Thompson, M. E., Fong, G. T., Hammond, D., Boudreau, C., Driezen, P., Hyland, A., et al. (2006). Methods of the international tobacco control (ITC) four country survey. *Tobacco Control*, *15*(Suppl 3), i12–18.

Tibshirani, R. J. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, Series B*, *58*, 267–288.

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. NewYork: Springer.

van der Vaart, A. W. (2000). *Asymptotic Statistics* (Vol. 3). Cambridge: Cambridge University Press.

Wu, C., & Luan, Y. (2003). Optimal calibration estimators under two-phase sampling. *Journal of Official Statistics*, *19*, 119–131.

Wu, C., & Rao, J. N. K. (2006). Pseudo empirical likelihood ratio confidence intervals for complex surveys. *The Canadian Journal of Statistics*, *34*, 359–375.

Wu, C., & Thompson, M. E. (2020). *Sampling Theory and Practice*. Cham: Springer.

Zhao, P., & Wu, C. (2019). Some theoretical and practical aspects of empirical likelihood methods for complex surveys. *International Statistical Review*, *87*, S239–S256.

Zhao, P., Haziza, D., & Wu, C. (2019). Sample empirical likelihood and the design-based oracle variable selection theory. Statistica Sinica, second revision submitted.

Zhao, P., Rao, J.N.K., & Wu, C. (2020). Empirical likelihood methods for public-use survey data. arXiv : 2005.12172.

Zhao, P., Ghosh, M., Rao, J. N. K., & Wu, C. (2020). Bayesian empirical likelihood inference with complex survey data. *Journal of the Royal Statistical Society, Series B*, *82*, 155–174.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.