Statistica Sinica Preprint No: SS-2019-0058	
Title	Sample Empirical Likelihood and the Design-based Oracle Variable Selection Theory
Manuscript ID	SS-2019-0058
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202019.0058
Complete List of Authors	Puying Zhao, David Haziza and Changbao Wu
Corresponding Author	Changbao Wu
E-mail	cbwu@uwaterloo.ca
Notice: Accepted version subject to English editing.	

Statistica Sinica

Sample Empirical Likelihood and the Design-based Oracle Variable Selection Theory

Puying Zhao[1], David Haziza[2] and Changbao Wu[3]

[1] *Yunnan University,* [2] *Université de Montréal and* [3] *University of Waterloo*

Abstract: The sample empirical likelihood approach provides a powerful tool for analysis of complex survey data. We present results of sample empirical likelihood for point estimation and linear or nonlinear hypothesis tests on finite population parameters defined through just-identified or over-identified estimating equation systems with smooth or non-differentiable estimating functions under general unequal probability sampling designs. We propose a penalized sample empirical likelihood for variable selection and establish its oracle property under the design-based framework. Practical implementations of the methods are also discussed. Finite sample performances of the proposed methods for quantile regression and variable selection are examined through simulation studies. An application using the survey dataset from the International Tobacco Control (ITC) Policy Evaluation Project is presented to demonstrate the effectiveness of the variable selection method for linear and quantile regression models.

Key words and phrases: Design-based variable selection theory, General hypothesis test, Non-differentiable estimating functions, Over-identified estimating equation system, Quantile regression analysis, Unequal probability sampling.

1. Introduction

Complex surveys are an important tool of data collection for many areas of scientific investigation. Survey data are widely used for official statistics, social science researches and population health studies. Design-based inferences are the predominant approach in official statistics for descriptive population parameters such as the population mean or quantiles. There has also been increased use of survey data for analytical purposes, such as exploring the relations among variables or building statistical models for estimation and prediction. The use of survey weights for analytical studies, however, is a subject of debate over the past three decades. One of the central concepts for valid model-based inferences using survey data is the ignorability of survey design features. Pfeffermann (1993) and Gelman (2007) contain stimulating discussions on the topic.

Design-based estimating equations approach has gained increased popularity among survey researchers and survey data users. The estimating functions are motivated by the inferential problems for the superpopulation model parameters θ, and the finite population parameters θ_N are defined as the solution to the so-called census estimating equations. Inferences are carried out through the survey weighted estimating equations. The survey weighted estimators $\hat{\theta}$ are typically design-consistent for the finite population parameters θ_N regardless of the model, and are also valid estimators for the model parameters θ if the model holds for the finite population (Godambe and Thompson, 1986). Design-based

variance estimators are also valid for the estimation of model parameters under the joint randomization of the superpopulation model and the survey sampling design (Binder and Roberts, 2009).

Empirical likelihood was first studied by Owen (1988) for independent data and has since become one of the fastest growing topics in statistics. Due to its nonparametric features, the method has been discussed extensively in the survey sampling literature for design-based inferences. The very first use of empirical likelihood method in surveys, however, is credited to Hartley and Rao (1968) under the "scale-load" approach. Chen and Qin (1993) presented the first formal use of the empirical likelihood for estimating the population mean under simple random sampling. For general unequal probability sampling designs, Chen and Sitter (1999) considered the pseudo empirical likelihood for complex survey designs with the main focus on point estimation for the population mean. Wu and Rao (2006) proposed pseudo empirical likelihood ratio confidence intervals which are applicable to a single parameter under arbitrary sampling designs, and Rao and Wu (2010) extended the method to multiple frame surveys.

Chen and Kim (2014) proposed the population empirical likelihood approach for parameters defined through estimating equations. They focused on Poisson sampling and conditional Poisson sampling and established an optimal property of the point estimator as well as the asymptotic chi-square distribution of the empirical likelihood ratio statistic with smooth estimating functions. Chen

and Kim (2014) also briefly introduced the sample empirical likelihood approach, which is a variation to the population empirical likelihood. The sample empirical likelihood estimator is algebraically equivalent to the nonparametric likelihood estimator introduced by Kim (2009). Berger and Torres (2016) and Oguz-Alper and Berger (2016) studied an empirical likelihood approach by incorporating a design-specific constraint such that the empirical likelihood ratio statistic for a scalar parameter or a subset of the vector parameters follows asymptotically a standard chisquare distribution. They considered settings where the parameters are defined through estimating equations system over-identified with calibration constraints. They illustrate their results for four commonly used unequal probability sampling designs. Non-smooth estimating equations are considered in Berger and Torres (2016). This approach is generalized for the multidimensional case by Oguz-Alper and Berger (2016). However, in Oguz-Alper and Berger (2016), the empirical likelihood test is based on profiling and differentiability. Berger (2016) discussed the method for the Rao-Hartley-Cochran sampling design and Berger (2018) addressed issues with nonresponse under cluster sampling. A key feature of the specific survey designs considered by Berger and his co-authors is that the design-based variance can be approximated without involving the second order inclusion probabilities. Standard chisquare limiting distribution does not hold for arbitrary sampling designs.

This paper provides a unified treatment on sample empirical likelihood ap-

proach to design-based survey data analysis. We consider the most general setting where the vector of finite population parameters is defined through a just-identified or over-identified census estimating equations system with smooth or non-differential estimating functions, with or without additional calibration constraints. The main theoretical results are established under the general setting with an arbitrary unequal probability sampling design. However, unknown joint-inclusion probabilities may be needed for testing, for the estimation of eigenvalues. Approximating these probabilities is often inevitable (e.g., Haziza et al., 2008) and requires assumptions about the design such as high entropy or negligible sampling fractions, as in Berger and Torres (2016) and Oguz-Alper and Berger (2016) (see Section 4 in the supplement for more details). Our paper contains four major methodological contributions to design-based survey data analyses: (i) The establishment of design-consistency and asymptotic normality of the maximum sample empirical likelihood estimator; (ii) The development of sample empirical likelihood ratio tests for a general linear or nonlinear hypothesis on finite population parameters; (iii) A rigorous treatment on parameters defined through non-differentiable estimating functions, with the general design-based results from (i) and (ii) covering advanced inference problems such as quantile regression analysis; and (iv) The penalized sample empirical likelihood method for design-based variable selection using complex survey data and the establishment of its oracle properties for parameters defined through general estimating

equations. The sample empirical likelihood formulation also brings a computational unification of design-based inferences for surveys and the mainstream applications of empirical likelihood (Owen, 1988; Qin and Lawless, 1994, 1995). Our asymptotic development uses the theory of empirical processes and extends existing methods for non-smooth estimation problems (Pakes and Pollard, 1989; Parente and Smith, 2011) from independent samples to complex survey data.

The rest of the paper is organized as follows. The general results of sample empirical likelihood on point estimation and linear or nonlinear hypothesis tests are presented in Section 2. Our proposed penalized sample empirical likelihood method for design-based variable selection and its oracle properties are given in Section 3. Issues with practical implementations of the methods are discussed in Section 4. Finite sample performances of the methods for quantile regression are examined through simulation studies reported in Section 5. An application using the survey dataset from the International Tobacco Control (ITC) Policy Evaluation Project is presented in Section 6 to demonstrate the effectiveness of the variable selection method. Some additional remarks are given in Section 7. Proofs of major results and further technical details, along with regularity conditions, computational details and additional simulation results, are reported in the Supplementary Materials.

2. Sample Empirical Likelihood Inference for Complex Surveys

We follow the conventional asymptotic framework for design-based infer-

ences. Suppose that we have a sequence of finite populations $\mathcal{U}_\nu = \{1, 2, \cdots, N_\nu\}$ indexed by ν. The population size $N_\nu \to \infty$ as $\nu \to \infty$. We drop the index ν for notational convenience and use $N \to \infty$ to represent the limiting process. Associated with each unit $i \in \mathcal{U} = \{1, 2, \cdots, N\}$ are values of survey variables (X_i, Y_i), where the Y_i represents the vector of study variables, the X_i denotes the vector of auxiliary variables. Let $\mathcal{F}_N = \{(X_1, Y_1), \cdots, (X_N, Y_N)\}$ be the set of values of all variables for the finite population. The finite population parameters of interest, denoted as $\theta_N \in \Theta$ where Θ is a compact subset of $\mathcal{R}^p$ and p is the dimension of θ_N, are defined as the solution to the census estimating equations

$$U_N(\theta) = \frac{1}{N} \sum_{i=1}^{N} g(X_i, Y_i, \theta) = 0, \tag{2.1}$$

where $g(X, Y, \theta)$ is an estimating function of dimension $r(\geq p)$. Most descriptive finite population parameters, such as means, proportions, distribution functions, quantiles and domain means, can be defined through (2.1). Moreover, statistical inferences for commonly encountered model parameters can be carried out through the finite population parameters defined through (2.1). Examples include linear regression models and generalized linear models.

We consider the general setting where the estimating equations system (2.1) is either just-identified (i.e., $r = p$) or over-identified (i.e., $r > p$), and the estimating function $g(X_i, Y_i, \theta)$ is either smooth or non-differentiable in θ. There are three scenarios for over-identified estimating equations systems. The first scenario is that all equations involve the parameters but there are more equations

than the parameters. A simple example is the Poisson distribution where the mean parameter also satisfies the equation for the variance. Another practically important example is the instrumental variable regression, where the number of equations r is the same as the number of instrumental variables and is often larger than the number of parameters p. See, for instance, Bowden and Turkington (1984). The second scenario often appears in survey sampling as the additional calibration constraints, which do not involve the parameters θ. The third scenario is a combination of the first two scenarios. The inclusion of calibration constraints to create an artificial "over-identified" system has been discussed by several authors, including Chen and Kim (2014), Berger and Torres (2016), and Oguz-Alper and Berger (2016), among others. The scenario can be handled by using the survey weighted estimating equations for the constrained maximization problem. Our theoretical results do not distinguish among specific scenarios and are developed under the general setting.

There are limited work on finite population parameters defined through non-differentiable estimating functions in the existing literature on survey data analysis. The work of Berger and Torres (2016) is based on non-differentiable estimations defining a scalar parameter. Oguz-Alper and Berger (2016) multi-dimensional test is based on profiling which can be easily implemented under differentiability, as in Qin and Lawless (1994). In Berger and Torres (2016) and Oguz-Alper and Berger (2016), differentiability is not needed for point estimation

or for testing the whole parameter θ_N. For non-differentiable $g(X, Y, \theta)$, exact solutions to $U_N(\theta) = 0$ may not exist. Under such scenarios we may replace (2.1) by $U_N(\theta) = a_N$ for some sequence $a_N = O(N^{-1})$. The introduction of a_N is for convenience in asymptotic developments involving non-differentiable estimating functions and has no practical implications. An important application involving non-differentiable estimating functions is quantile regression analysis using survey data with $g(X, Y, \theta) = X\{I(Y < X^{\mathrm{T}}\theta) - \tau\}$, where $\tau \in (0, 1)$ and $I(\cdot)$ is the indicator function. We aim to develop a unified theory to cover both smooth and non-differentiable estimating functions.

2.1. Point estimation and asymptotic properties

Let $\mathcal{S}$ be the set of sampled units selected by a probability sampling design, with first and second order inclusion probabilities $\pi_i = \mathrm{P}(i \in \mathcal{S})$ and $\pi_{ij} = \mathrm{P}(i, j \in \mathcal{S})$. Let n be the realized sample size which could be random under certain sampling designs such as Poisson sampling. Let $n_B = \mathrm{E}(n \mid \mathcal{F}_N)$ be the expected sample size. Issues with unit or item nonresponses are not considered in the current paper. The survey sample dataset is denote by $\{(X_i, Y_i), i \in \mathcal{S}\}$.

Let $(p_1, \cdots, p_n)$ be the discrete probability measure assigned to the n sampled units. Let $g_i(\theta) = g(X_i, Y_i, \theta)$. The sample empirical likelihood (SEL) function for the given θ is defined as

$$L_n(\theta) = \sup \left\{ \prod_{i \in \mathcal{S}} p_i \;\Big|\; p_i \geq 0, \sum_{i \in \mathcal{S}} p_i = 1, \sum_{i \in \mathcal{S}} p_i \left[\pi_i^{-1} g_i(\theta)\right] = 0 \right\}.$$

It should be noted that the definition of $L_n(\theta)$ follows the standard formulation of

Qin and Lawless (1994) on empirical likelihood and estimating equations with one simple modification: the basic design weight π_i^{-1} is treated as an intrinsic part of the estimating function $g_i(\theta)$. This is sufficient to obtain design-consistent point estimators but hypothesis tests will require design-based variance estimation for a general sampling design. The formulation brings a computational unification between the sample empirical likelihood in survey sampling and empirical likelihood methods in other areas, and permits advanced asymptotic development presented in this paper. It should be noted that our sample empirical likelihood formulation follows Chen and Kim (2014) but equivalent formulations are also used in Berger and Torres (2012, 2014, 2016) and Oguz-Alper and Berger (2016). Chen and Kim (2014) relies on Poisson sampling. The use of π_i^{-1} as part of the constraints also appeared in Kim's (2009) formulation which is however different.

By the standard derivation of empirical likelihood (Qin and Lawless, 1994), the sample empirical likelihood function for any given θ, is given by $L_n(\theta) = \prod_{i\in\mathcal{S}} \hat{p}_i(\theta)$, where $\hat{p}_i(\theta) = \{n[1 + \lambda^{\mathrm{T}}\pi_i^{-1}g_i(\theta)]\}^{-1}$ and the Lagrange multiplier $\lambda = \lambda(\theta)$ is the solution to

$$\sum_{i\in\mathcal{S}} \frac{\pi_i^{-1}g_i(\theta)}{1 + \lambda^{\mathrm{T}}\pi_i^{-1}g_i(\theta)} = 0\,. \tag{2.2}$$

We have $\log\{L_n(\theta)\} = -l_n(\theta, \lambda) - n\log(n)$, where

$$l_n(\theta, \lambda) = \sum_{i\in\mathcal{S}} \log\{1 + \lambda^{\mathrm{T}}\pi_i^{-1}g_i(\theta)\}. \tag{2.3}$$

Finding the solution $\lambda = \lambda(\theta)$ to (2.2) is a dual problem of maximizing $l_n(\theta, \lambda)$

with respect to λ for the given θ. The maximum sample empirical likelihood estimator $\hat{\theta}_{SEL}$ for θ_N is given by

$$\hat{\theta}_{SEL} = \arg\min_{\theta\in\Theta} \sup_{\lambda\in\hat{\Lambda}_n(\theta)} l_n(\theta,\lambda)\,,$$

where $\hat{\Lambda}_n(\theta) = \{\lambda \mid \lambda^{\mathrm{T}}\pi_i^{-1}g_i(\theta) > -1, i \in \mathcal{S}\}$ for the given θ.

Our first major theoretical result is on design-consistency and asymptotic normality of the maximum sample empirical likelihood estimator $\hat{\theta}_{SEL}$. The required regularity conditions C1-C6 are specified in the Supplementary Materials. Proofs of the results under the general setting involve extending the modern empirical process theory (e.g., Pakes and Pollard, 1989; van der Vaart and Wellner, 1996; Chen et al., 2003) for independent data to dependent complex survey data. Let $\|A\| = \{\mathrm{trace}(A^{\mathrm{T}}A)\}^{1/2}$ for any matrix or vector A.

Theorem 1. *Suppose that Conditions C1-C6 given in the Supplementary Materials hold. We have that*

(i) The maximum sample empirical likelihood estimator $\hat{\theta}_{SEL}$ is design-consistent for θ_N, i.e., for any $\epsilon > 0$,

$$\lim_{N\to\infty} \mathrm{P}\{\|\hat{\theta}_{SEL} - \theta_N\| > \epsilon \mid \mathcal{F}_N\} = 0\,.$$

(ii) The estimator $\hat{\theta}_{SEL}$ is asymptotically normally distributed with mean θ_N and variance-covariance matrix

$$V_1 = (\Gamma^{\mathrm{T}}W^{-1}\Gamma)^{-1}\Gamma^{\mathrm{T}}W^{-1}\Omega W^{-1}\Gamma(\Gamma^{\mathrm{T}}W^{-1}\Gamma)^{-1}\,,$$

where $\Gamma = \Gamma(\theta_N)$, $\Gamma(\theta) = \partial U(\theta)/\partial\theta$ *with* $U(\theta)$ *as the limiting function of* $U_N(\theta)$ *defined in Condition C2,* $W = n_B N^{-2} \sum_{i=1}^{N} \pi_i^{-1}[g_i(\theta_N)][g_i(\theta_N)]^{\mathrm{T}}$, $\Omega = \mathrm{Var}\big[\hat{U}_N(\theta_N) \mid \mathcal{F}_N\big]$, *and* $\hat{U}_N(\theta) = N^{-1} \sum_{i\in\mathcal{S}} \pi_i^{-1} g_i(\theta)$.

It should be noted that the factor $n_B N^{-2}$ in the definition of W is not needed for defining V_1 since it all cancels out. The design-based variance Ω may depend on joint inclusion probabilities. Applications of Theorem 1 and other general results presented in this section require estimation of quantities such as Γ, W and Ω, which are discussed in Section 4.

Corollary 1. *Under the same regularity conditions for Theorem 1, the asymptotic variance-covariance matrix* V_1 *for* $\hat{\theta}_{SEL}$ *is simplified for the following two special cases:*

(i) If $r = p$, *then* V_1 *reduces to* $V_2 = \Gamma^{-1}\Omega(\Gamma^{\mathrm{T}})^{-1}$.

(ii) Under single-stage PPS sampling with replacement or single-stage PPS sampling without replacement with negligible sampling fractions, the variance-covariance matrix V_1 *reduces to* $V_3 = (n_B\Gamma^{\mathrm{T}}W^{-1}\Gamma)^{-1}$.

Theorem 1 and and Corollary 1 are important for our subsequent analysis and play a key role in establishing limiting distributions of sample empirical likelihood ratio test statistics presented in the next section.

2.2. Empirical likelihood ratio tests for general hypotheses on θ_N

We first present the result on the asymptotic distribution of the sample em-

pirical likelihood ratio statistic for θ_N under the general setting and an arbitrary sampling design. The sample empirical likelihood ratio statistic for θ_N is defined as

$$T_n(\theta) = -2\{l_n(\hat{\theta}_{SEL}, \hat{\lambda}_{SEL}) - l_n(\theta, \lambda)\}\,, \tag{2.4}$$

where $\hat{\lambda}_{SEL} = \hat{\lambda}(\hat{\theta}_{SEL})$ and $\hat{\lambda}(\theta) = \arg\sup_{\lambda \in \hat{\Lambda}_n(\theta)} l_n(\theta, \lambda)$. We have the following general result on the asymptotic distribution of $T_n(\theta_N)$.

Theorem 2. *Suppose that Conditions C1-C6 given in the Supplementary Materials hold. Then, as $N \to \infty$, $T_n(\theta_N)$ converges in distribution to $Q^{\mathrm{T}}\Delta Q$ when θ_N is the true value of the vector parameter, where Q follows the standard multivariate normal distribution $N(0, I_r)$ and*

$$\Delta = n_B \Omega^{1/2} W^{-1} \Gamma (\Gamma^{\mathrm{T}} W^{-1} \Gamma)^{-1} \Gamma^{\mathrm{T}} W^{-1} \Omega^{1/2}\,.$$

The asymptotic distribution of $T_n(\theta_N)$ can be alternatively represented by $\sum_{j=1}^{p} \delta_j \chi_j^2$, where χ_j^2, $j = 1, \cdots, p$ are independent random variables, all following the same distribution as χ^2 with one degree of freedom, and δ_j, $j = 1, \cdots, p$ are the non-zero eigenvalues of the $r \times r$ matrix Δ. For the special case $r = p = 1$, the parameter θ becomes a scalar and Δ reduces to a constant $a = \mathrm{Var}\{\sum_{i \in \mathcal{S}} \pi_i^{-1} g_i(\theta_N) \mid \mathcal{F}_N\} / \sum_{i=1}^{N} \pi_i^{-1} [g_i(\theta_N)]^2$. In general, the test statistic $T_n(\theta_N)$ follows a scaled χ^2 distribution with one degree of freedom when $p = 1$, which is similar to the main result presented in Wu and Rao (2006).

Corollary 2. *Suppose that Conditions C1-C6 given in the Supplementary Ma-*

terials hold. Under single-stage PPS sampling with replacement or single-stage PPS sampling without replacement with negligible sampling fractions, the sample empirical likelihood ratio statistic $T_n(\theta_N)$ converges in distribution to a χ^2 random variable with p degrees of freedom as $N \to \infty$, where p is the dimension of θ_N.

Corollary 2 can also be found in Oguz-Alper and Berger (2016) and in Berger and Torres (2016) (when $p = 1$), where differentiability is not needed to establish the result. For the sampling designs described in Corollary 2, the $(1-\alpha)100\%$ confidence region for θ_N can be constructed as

$$C_\alpha = \left\{\theta \mid -2\{l_n(\hat{\theta}_{SEL}, \hat{\lambda}_{SEL}) - l_n(\theta, \lambda)\} \leq \chi^2_{1-\alpha}(p)\right\},$$

where $\chi^2_{1-\alpha}(p)$ is the $1-\alpha$ quantile of the χ^2 distribution with p degrees of freedom. For a general sampling design, the value $\chi^2_{1-\alpha}(p)$ needs to be replaced by the $1-\alpha$ quantile from $\sum_{j=1}^{p} \hat{\delta}_j \chi^2_j$, where the $\hat{\delta}_j$, $j = 1, \cdots, p$ are the estimated non-zero eigenvalues of Δ given in Theorem 2.

We now consider sample empirical likelihood ratio tests for a general hypothesis H_0: $\Phi(\theta_N) = 0$ against a suitable alternative, where $\Phi(\theta)$ has k $(\leq p)$ smooth components and $\Phi(\theta) = 0$ imposes k constraints on θ, either linear or nonlinear. Let $\Theta^* = \left\{\theta \mid \theta \in \Theta \text{ and } \Phi(\theta) = 0\right\}$ be the restricted parameter space under H_0. The restricted maximum sample empirical likelihood estimator of θ under H_0 is defined as $\hat{\theta}^*_{SEL} = \arg\min_{\theta \in \Theta^*} \sup_{\lambda \in \hat{\Lambda}_n(\theta)} l_n(\theta, \lambda)$.

Let $\hat{\lambda}^*_{SEL} = \arg\sup_{\lambda \in \hat{\Lambda}_n(\hat{\theta}^*_{SEL})} l_n(\hat{\theta}^*_{SEL}, \lambda)$. The sample empirical likelihood ratio statistic for testing H_0: $\Phi(\theta_N) = 0$ against a suitable alternative is defined as

$$T_n(\theta_N \mid H_0) = -2\{l_n(\hat{\theta}_{SEL}, \hat{\lambda}_{SEL}) - l_n(\hat{\theta}^*_{SEL}, \hat{\lambda}^*_{SEL})\}.$$

Let $\Psi(\theta) = \partial\Phi(\theta)/\partial\theta$, which is a $k \times p$ matrix. We assume that $\Psi(\theta)$ has full rank k.

Theorem 3. *Suppose that Conditions C1-C6 given in the Supplementary Materials and the null hypothesis H_0: $\Phi(\theta_N) = 0$ hold. As $N \to \infty$, we have that*

(i) The restricted maximum sample empirical likelihood estimator $\hat{\theta}^*_{SEL}$ has asymptotic variance-covariance matrix given by

$$V^* = P_1^* \Gamma^{\mathrm{T}} W^{-1} \Omega W^{-1} \Gamma P_1^*,$$

where $P_1^* = \Sigma - \Sigma\Psi^{\mathrm{T}}(\Psi\Sigma\Psi^{\mathrm{T}})^{-1}\Psi\Sigma$, $\Sigma = (\Gamma^{\mathrm{T}} W^{-1}\Gamma)^{-1}$ and $\Psi = \Psi(\theta_N)$.

(ii) The sample empirical likelihood ratio statistic $T_n(\theta_N \mid H_0)$ converges in distribution to $Q^{\mathrm{T}}\Delta^* Q$, where $Q \sim N(0, I_r)$ and

$$\Delta^* = n_B \Omega^{1/2} W^{-1}\Gamma(\Sigma - P_1^*)\Gamma^{\mathrm{T}} W^{-1}\Omega^{1/2}.$$

We notice that Δ^* has the same structure as Δ from Theorem 2, with the central piece Σ in Δ replaced by $\Sigma - P_1^* = \Sigma\Psi^{\mathrm{T}}(\Psi\Sigma\Psi^{\mathrm{T}})^{-1}\Psi\Sigma$ for Δ^*. Under general settings the distribution of $Q^{\mathrm{T}}\Delta^* Q$ is a weighted χ^2 involving eigenvalues of Δ^*. If $r = p$, $\Delta^* = n_B\Omega^{1/2}(\Gamma^{\mathrm{T}})^{-1}\Psi^{\mathrm{T}}(\Psi\Sigma\Psi^{\mathrm{T}})^{-1}\Psi\Gamma^{-1}\Omega^{1/2}$. Theorem 2 of Oguz-Alper and Berger (2016) presented a result similar to Part (ii) under the setting

that $\hat{U}_N(\theta)$ is differentiable, the over-identified system is specified by calibration constraints and the $\Phi(\theta)$ defines a sub-parameter. The result in Part (ii) is simplified for single-stage PPS sampling designs.

Corollary 3. *Suppose that Conditions C1-C6 given in the Supplementary Materials hold. Under single-stage PPS sampling with replacement or single-stage PPS sampling without replacement with negligible sampling fractions, the sample empirical likelihood ratio statistic* $T_n(\theta_N \mid H_0)$ *converges in distribution to a* χ^2 *random variable with* k *degrees of freedom as* $N \to \infty$.

A similar simplified result was presented in Oguz-Alper and Berger (2016) under the setting used in their paper. There are two practically important applications of the general results presented in Theorem 3 and Corollary 3. The first is for testing a linear hypothesis on θ_N in the form of H_0: $A\theta_N = 0$, where A is a known $k \times p$ matrix. In this case we have $\Psi(\theta) = \partial\Phi(\theta)/\partial\theta = A$. Let $\theta_N = (\theta_{N1}^{\mathrm{T}}, \theta_{N2}^{\mathrm{T}})^{\mathrm{T}}$ be a partition of the parameters. The most commonly used linear hypothesis for model building is H_0: $\theta_{N2} = 0$, corresponding to a simple form of A.

The second application is to construct confidence intervals or regions in the presence of nuisance parameters. This is the topic discussed by Oguz-Alper and Berger (2016) using a profile empirical likelihood method. Suppose that θ_{N1} is the vector of the parameters of interest and θ_{N2} is treated as the vector of nuisance parameters. Let $\theta = (\theta_1^{\mathrm{T}}, \theta_2^{\mathrm{T}})^{\mathrm{T}}$ correspond to the same partition as θ_N

with dimension k for θ_{N1} and θ_1. For single-stage PPS sampling designs, the $(1-\alpha)100\%$ confidence region for θ_{N1} can be constructed as

$$C_\alpha^* = \left\{\theta_1 \mid -2\{l_n(\hat{\theta}_{SEL}, \hat{\lambda}_{SEL}) - l_n[\tilde{\theta}(\theta_1), \tilde{\lambda}]\} \leq \chi^2_{1-\alpha}(k)\right\}, \tag{2.5}$$

where $\tilde{\theta}(\theta_1) = (\theta_1^{\mathrm{T}}, \hat{\theta}_2(\theta_1)^{\mathrm{T}})^{\mathrm{T}}$, $\hat{\theta}_2(\theta_1) = \arg\min_{\theta_2} l_n(\theta, \lambda)$ for the given θ_1, and $\tilde{\lambda} = \arg\sup_{\lambda \in \hat{\Lambda}_n(\tilde{\theta}(\theta_1))} l_n(\tilde{\theta}(\theta_1), \lambda)$. For general sampling designs, the cut-off point $\chi^2_{1-\alpha}(k)$ needs to be replaced by the estimated $1-\alpha$ quantile from the weighted χ^2 distribution given in Theorem 3.

3. Design-based Variable Selection and Its Oracle Property

Complex surveys often collect information on a large number of variables. Some of those variables measure basic characteristics of the units and some are specifically designed for broad scientific objectives. Section 6 presents an example from the ITC Project where many variables related to demographical, psychosocial, behavioural and health aspects of the units are measured for the survey data. The initial stage for model building requires identification and selection of important factors for several responses on addiction and quitting behaviours. Variable selection for complex survey data is an important topic that has not been fully addressed in the existing literature.

Under the non-survey context, the basic setting for variable selection is to identify variables in a regression model with the coefficients being zero. For finite population regression coefficients θ_N defined as the solution to the census estimating equations, the components of θ_N are usually not exactly equal to zero

even if the corresponding superpopulation parameters are zero. The usual root-n order implies that $\theta_N = O(N^{-1/2})$ if the model parameters are zero and the model holds for the finite population. We consider practical scenarios where N is very large and certain components of θ_N can be treated as zero, corresponding to the zero coefficients in the superpopulation model.

We consider the SCAD penalty $p_{\tau_n}(\cdot)$ proposed by Fan and Li (2001) with a tuning parameter τ_n to be selected by a data-driven method. To estimate θ_N and identify its nonzero components, we propose to use the following penalized sample empirical likelihood function

$$l_{\tau_n}(\theta) = \sum_{i\in\mathcal{S}} \log\{1 + \lambda^{\mathrm{T}}\pi_i^{-1}g_i(\theta)\} + n\sum_{j=1}^{p} p_{\tau_n}(|\theta_j|)\,, \tag{3.1}$$

where λ solves the equation given by (2.2) with the given θ. The penalty function $p_{\tau_n}(\cdot)$ satisfies $p_{\tau_n}(0) = 0$ with its first order derivative given by

$$p'_{\tau}(\theta) = \tau\left\{\mathrm{I}(\theta \le \tau) + \frac{(a\tau - \theta)_+}{(a-1)\tau}\mathrm{I}(\theta > \tau)\right\},$$

where $a > 2$. Fan and Li (2001) recommended to use $a = 3.7$ for most applications.

Let $\theta_{N[j]}$ be the jth component of θ_N and let $\mathcal{A} = \{j \mid 1 \le j \le p \text{ and } \theta_{N[j]} \neq 0\}$ be the index set of the nonzero components. For asymptotic development, two conditions are assumed (Fan and Li, 2001) for the penalty function and the tuning parameter τ_n. See Supplementary Materials for further detail.

Without loss of generality, we assume that $\theta_N = (\theta_{N1}^{\mathrm{T}}, \theta_{N2}^{\mathrm{T}})^{\mathrm{T}}$, where θ_{N1}

consists of the d non-zero components and $\theta_{N2} = 0$. Let $\hat{\theta}_{PSEL} = (\hat{\theta}_{P1}^{\mathrm{T}}, \hat{\theta}_{P2}^{\mathrm{T}})^{\mathrm{T}}$ be the penalized maximum sample empirical likelihood estimator of θ_N, which is the minimizer of (3.1). Using the partition of θ_N, we decompose the variance-covariance matrices V_j $(j = 1, 2, 3)$ and Σ defined in Theorem 3 and Corollary 3 into the following block matrices

$$V_j = \begin{pmatrix} V_{j11} & V_{j12} \\ V_{j21} & V_{j22} \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

The asymptotic properties of the penalized estimator $\hat{\theta}_{PSEL} = (\hat{\theta}_{P1}^{\mathrm{T}}, \hat{\theta}_{P2}^{\mathrm{T}})^{\mathrm{T}}$ are summarized in the following theorem and corollary.

Theorem 4. *Suppose that Conditions C1-C8 given in the Supplementary Materials hold, and that $p = \dim(\theta_N)$ is finite. Then as $N \to \infty$,*

(i) The penalized maximum sample empirical likelihood estimator $\hat{\theta}_{P2}$ for the zero-components satisfies $\mathrm{P}(\hat{\theta}_{P2} = 0 \mid \mathcal{F}_N) \to 1$.

(ii) The estimator $\hat{\theta}_{P1}$ for the non-zero components is asymptotically normal with mean θ_{N1} and variance-covariance matrix

$$V_{1\mathcal{A}} = V_{111} - V_{112}\Sigma_{22}^{-1}\Sigma_{21} - \Sigma_{12}\Sigma_{22}^{-1}V_{121} + \Sigma_{12}\Sigma_{22}^{-1}V_{122}\Sigma_{22}^{-1}\Sigma_{21}\,.$$

Part (i) of Theorem 4 is the oracle property of design-based variable selection through the penalized sample empirical likelihood method. The most crucial difference between our proposed approach and the standard approach for independent data is the use of survey weighted constraints in defining the sample

empirical likelihood function, which is the first part of $l_{\tau_n}(\theta)$ given by (3.1). It ensures design-consistency of the unpenalized point estimator, which is the foundation for the penalized approach for variable selection under the design-based framework.

Corollary 4. *Under the same regularity conditions for Theorem 4, the asymptotic variance-covariance matrix $V_{1\mathcal{A}}$ for $\hat{\theta}_{P1}$ is simplified for the following two special cases:*

(i) If $r = p$, then V_1 reduces to $V_2 = \Gamma^{-1}\Omega(\Gamma^{\mathrm{T}})^{-1}$ and $V_{1\mathcal{A}}$ is given by

$$V_{2\mathcal{A}} = V_{211} - V_{212}\Sigma_{22}^{-1}\Sigma_{21} - \Sigma_{12}\Sigma_{22}^{-1}V_{221} + \Sigma_{12}\Sigma_{22}^{-1}V_{222}\Sigma_{22}^{-1}\Sigma_{21}.$$

(ii) Under single-stage PPS sampling with replacement or single-stage PPS sampling without replacement with negligible sampling fractions, the asymptotic variance-covariance matrix $V_{1\mathcal{A}}$ reduces to $V_{3\mathcal{A}} = V_{311} - V_{312}V_{322}^{-1}V_{321}$.

Our proposed penalized sample empirical likelihood also provides more efficient estimation for the nonzero components of the parameters in terms of smaller asymptotic variances as shown in Corollary 4. By using a nonconvex penalty function such as the SCAD, the penalized maximum sample empirical likelihood estimator of θ_N is asymptotically equivalent to the "unpenalized" maximum sample empirical likelihood estimator of Section 2 on the restricted parameter space $\Theta^* = \{\theta \mid \theta \in \Theta \text{ and } \theta_{N2} = 0\}$.

The penalized sample empirical likelihood can be further used as a general

tool to conduct hypotheses tests on the $d \times 1$ non-zero components of θ_N as subsequent steps to variable selection. Specifically, we consider the problem of testing linear hypotheses

$$H_0 : B\theta_{N1} = 0\,, \quad H_1 : B\theta_{N1} \neq 0\,, \tag{3.2}$$

where the known $q \times d$ matrix B satisfies $BB^{\mathrm{T}} = I_q$ with fixed q. We assume that $q < d$, i.e., the number of constraints in H_0: $B\theta_{N1} = 0$ is smaller than the number of parameters in θ_{N1}, which excludes H_0: $\theta_{N1} = 0$ from consideration. In a model-based parametric likelihood framework, Fan and Peng (2004) studied a similar type of hypothesis testing when the number of parameters is diverging with the sample size. The problem (3.2) includes hypotheses for individual and multiple components of θ_{N1} as special cases. The most common hypothesis is: $H_0 : \theta_{N1j} = 0\,, H_1 : \theta_{N1j} \neq 0\,$, where θ_{N1j} denotes the jth coordinate of θ_{N1}. The penalized sample empirical likelihood ratio function is computed as

$$T_{\tau_n}(\theta_{N1} \mid H_0) = -2\{l_{\tau_n}(\hat{\theta}_{PSEL}) - \min_{\theta : B\theta_1 = 0} l_{\tau_n}(\theta)\}\,, \tag{3.3}$$

where $\theta = (\theta_1^{\mathrm{T}}, \theta_2^{\mathrm{T}})^{\mathrm{T}}$ follows the same partition as $\theta_N = (\theta_{N1}^{\mathrm{T}}, \theta_{N2}^{\mathrm{T}})^{\mathrm{T}}$.

Theorem 5. *Suppose that Conditions C1-C8 given in the Supplementary Materials hold. Then, under single-stage PPS sampling with replacement or single-stage PPS sampling without replacement with negligible sampling fractions, the $T_{\tau_n}(\theta_{N1} \mid H_0)$ converges in distribution to a χ^2 random variable with q degrees of freedom.*

As a direct consequence, a $(1-\alpha)100\%$ confidence region for $\beta_N = B\theta_{N1}$ can be constructed as

$$C_\alpha^{[1]} = \left\{\beta \mid -2\{l_{\tau_n}(\hat{\theta}_{PSEL}) - \min_{\theta:B\theta_1=\beta} l_{\tau_n}(\theta)\} \leq \chi^2_{1-\alpha}(q)\right\}.$$

4. Practical Implementations

The asymptotic variance of $\hat{\theta}_{SEL}$ and the asymptotic distributions of the sample empirical likelihood ratio statistics presented in Sections 2 and 3 are derived under a general sampling design for smooth and non-differentiable estimating functions. Practical implementations of the methods require estimation of three major components: W, Γ and Ω, which further leads to the estimation of quantities such as V_1, Δ and Δ^*.

The first quantity $W = n_B N^{-2} \sum_{i=1}^N \pi_i^{-1}[g_i(\theta_N)][g_i(\theta_N)]^{\mathrm{T}}$ can be consistently estimated by $\hat{W} = n_B N^{-2} \sum_{i\in\mathcal{S}} \pi_i^{-2}[g_i(\hat{\theta}_{SEL})][g_i(\hat{\theta}_{SEL})]^{\mathrm{T}}$. As noted in Section 2, the factor $n_B N^{-2}$ is included for theoretical purposes and is not required for computation.

The second quantity is $\Gamma = \Gamma(\theta_N)$, where $\Gamma(\theta) = \partial U(\theta)/\partial\theta$, and $U(\theta)$ is the limiting function of $U_N(\theta) = N^{-1}\sum_{i=1}^N g_i(\theta)$. For smooth estimating functions with differentiable $g_i(\theta)$, we can use a simple plug-in estimator $\hat{\Gamma} = \hat{\Gamma}(\hat{\theta}_{SEL})$, where $\hat{\Gamma}(\theta) = \partial\hat{U}_N(\theta)/\partial\theta = N^{-1}\sum_{i\in\mathcal{S}} \pi_i^{-1}\partial g_i(\theta)/\partial\theta$. For nondifferentiable estimating functions, the estimation of Γ requires additional effort. We provide details for the quantile regression models used for the simulation studies reported in Section 5, where $g(X, Y, \theta) = X\{\mathrm{I}(Y < X^{\mathrm{T}}\theta) - \gamma\}$ for a prespecified $\gamma \in (0, 1)$.

Let $f(y|X)$ and $F(y|X)$ be, respectively, the conditional pdf and cdf of Y given X under the superpopulation model for (X, Y). It follows that $U_N(\theta) \to U(\theta)$ where

$$U(\theta) = E\big\{g(X, Y, \theta)\big\} = E\big[X\big\{\mathrm{P}(Y < X^{\mathrm{T}}\theta|X) - \gamma\big\}\big] = E\big[X\big\{F(X^{\mathrm{T}}\theta|X) - \gamma\big\}\big] .$$

It further leads to $\Gamma(\theta) = \partial U(\theta)/\partial\theta = E[f(X^{\mathrm{T}}\theta|X)XX^{\mathrm{T}}]$. Let $K(\cdot)$ be a kernel function. The quantity $\Gamma(\theta)$ with the given θ can be estimated by the survey weighted estimator

$$\hat{\Gamma}(\theta) = \frac{1}{Nh}\sum_{i\in\mathcal{S}} \pi_i^{-1} K\{(Y_i - X_i^{\mathrm{T}}\theta)/h\} X_i X_i^{\mathrm{T}} ,$$

where h is the bandwidth for kernel density estimation.

The estimation of the third quantity $\Omega = \mathrm{Var}[\hat{U}_N(\theta_N) \mid \mathcal{F}_N]$ amounts to design-based variance estimation for the Horvitz-Thompson estimator. This is one of the major topics in survey sampling and is not unique to the sample empirical likelihood methods developed in this paper. For single-stage PPS sampling without replacement with small sampling fractions, the results presented in Sections 2 and 3 do not require the estimation of Ω. We provide details in the Supplementary Materials for three other commonly encountered sampling designs in survey practice. Each of these designs will further be examined in the simulation studies.

5. Simulation Studies

We presents results from several simulation studies on the finite sample performances of the proposed methods for point estimation, hypothesis tests and variable selection. We focused on quantile regression (QR) models where the estimating functions are non-differentiable. The topic on QR itself has attracted increased attention in recent years on alternative regression modelling techniques. Results on point estimation and hypothesis tests are reported in this section. Results on variable selection for quantile regression models and results from another simulation on linear regression models are reported in the Supplementary Materials.

5.1. Basic settings and sampling designs

We considered design-based inferences where the finite population was generated from a superpopulation and was fixed for repeated simulation samples. We considered four sampling designs: (I) Single-stage PPS sampling without replacement with negligible sampling fractions; (II) Single-stage PPS sampling without replacement with non-negligible sampling fractions; (III) Stratified PPS sampling; (IV) Two-stage cluster sampling with self-weighting designs. Details on the finite population size, sample sizes and the four sampling designs are given in Section 6 of the Supplementary Document.

5.2. Point estimation and hypothesis tests

Our first simulation study investigated the design-based performances of the sample empirical likelihood estimator and the sample empirical likelihood ratio

tests for quantile regression models. The finite population was generated from the model

$$Y = \theta_0 + Z_1\theta_1 + Z_2\theta_2 + \sigma(Z_1, Z_2)(\varepsilon - Q_\varepsilon(\gamma)), \tag{5.1}$$

where $Q_\varepsilon(\gamma)$ is the γth quantile of ε. The scale factor $\sigma(Z_1, Z_2)$ allows for the presence of conditional heteroscedasticity. The true values of the model parameters were set as $(\theta_0, \theta_1, \theta_2) = (0.5, 1, 1)$, with the regressors $Z_1 \sim N(0, 1)$ and $Z_2 \sim \chi^2(3)$. We considered $\sigma(Z_1, Z_2) = 1$ and $\sigma(Z_1, Z_2) = 1 + Z_2$ to explore the effect of conditional heteroscedasticity. For the error term ε, we considered three scenarios: (i) $\varepsilon \sim N(0, 1)$; (ii) $\varepsilon \sim \chi^2(3)$; and (iii) $\varepsilon \sim t(3)$. Our simulation examined three quantile regression models corresponding to $\tau = 0.25$, 0.50 and 0.75.

Let $\theta = (\theta_0, \theta_1, \theta_2)^{\mathrm{T}}$, $X = (1, Z_1, Z_2)^{\mathrm{T}}$, and $X_i = (1, Z_{1i}, Z_{2i})^{\mathrm{T}}$, $i = 1, \cdots, N$. The finite population parameters $\theta_N(\gamma) = (\theta_{N0}(\gamma), \theta_{N1}(\gamma), \theta_{N2}(\gamma))^{\mathrm{T}}$ under the quantile regression model are defined through the census estimating equations $\sum_{i=1}^{N} g(X_i, Y_i, \theta_N(\gamma)) = 0$, where $g(X, Y, \theta) = X\{\mathrm{I}(Y < X^{\mathrm{T}}\theta) - \gamma\}$. Under the model (5.1) with the shifted $Q_\varepsilon(\gamma)$ for the error term, the true values of $\theta_N(\gamma) = (\theta_{N0}(\gamma), \theta_{N1}(\gamma), \theta_{N2}(\gamma))^{\mathrm{T}}$ for the finite population were essentially the same as the superpopulation model parameters $\theta = (0.5, 1, 1)^{\mathrm{T}}$. For hypothesis tests, we considered testing H_0: $\theta_{N1}(\gamma) = 1.0$ versus H_1: $\theta_{N1}(\gamma) = b$ for $b \in \{0.5, 0.75, 1.0, 1.25, 1.5\}$ to examine the size and the power of the sample empirical likelihood ratio test.

For each simulated sample, the maximum sample empirical likelihood estimator $\hat{\theta}_{SEL}$ was computed, and the three unknown quantities W, Γ and Ω were estimated using the methods described in Section 4. We used the Gaussian kernel function with bandwidth $h = n^{-1/3}$ for estimating Γ. The size and the power of the test were reported in Table 1 under the heteroscedastic error terms with $\sigma(Z_1, Z_2) = 1 + Z_2$. The simulated relative bias (Bias) and root mean squared error (RMS) of the point estimators and additional simulation results under the homogeneous structure $\sigma(Z_1, Z_2) = 1$ were included in the Supplementary Materials. The simulation results were based on $B = 2000$ repeated simulation samples and can be summarized as follows.

(a) Point estimation: The relative biases of the maximum sample empirical likelihood estimators are uniformly small ($< 3\%$) for all scenarios considered, including skewed error distributions, heteroscedasticity and different sampling designs. The values of RMS are similar to each other across the four different sampling designs but are smaller under homogeneous error terms or symmetric error distributions.

(b) Hypothesis test: The sizes of the test, corresponding to $b = 1.00$, are close to the nominal value 0.05 for the vast majority of cases included in the simulation. There are a few cases, mostly under the skewed error distribution $\chi^2(3)$, where the sizes of the test are slightly over the target (around 0.07). The power of the test (with $b \neq 1.00$) has demonstrated the effectiveness of the test for all scenarios, and the test is more powerful under the homogeneous structure $\sigma(Z_1, Z_2) = 1$.

This is consistent to general observations from other studies that the presence of heteroscedasticity has impact on the performance of tests for quantile regression models.

6. An Application to the ITC Survey Data

The International Tobacco Control Policy Evaluation Project (The ITC Project) conducts longitudinal surveys to measure the effectiveness of national-level tobacco control policies in more than 20 countries which signed and ratified the Framework Convention on Tobacco Control (FCTC). The ITC project first started in four countries: Canada, USA, Australia and the UK. The first wave ITC Four Country Survey used a stratified sampling design and conducted telephone interviews of over 2000 adult smokers in each of the four countries. The initial group of respondents was followed in subsequent waves and a new cross-sectional replenishment sample was added at each wave to make up for the reduced size of the longitudinal sample due to attrition. In wave 8, respondents were given options to complete the survey either through telephone interviews or by self-administered internet surveys with a user-specific link to the questionnaire pages. The ITC survey questionnaires cover a wide range of measures on demographic variables, smoking behaviour, warning labels, advertising and promotion, light/mild brand descriptors, taxation and purchase behaviour, stop-smoking medications and alternative nicotine products, cessation and quitting behaviour as well as key psychosocial variables. Thompson et al. (2006) contain

further details on the ITC Four Country Survey.

One of the important research problems in tobacco control is to model the relation between smoking addiction and factors such as those included in the ITC survey questionnaires. Due to the large number of potential variables available from the data file, variable selection techniques for the initial model building become highly valuable. In this section we apply the proposed sample empirical likelihood method to build a model for the response variable Y: Cigarettes Per Day, which is a common measure of the degree of addiction for smokers. We use the data set from the ITC Four Country wave 8 survey which contains $n = 901$ smokers from Canada. We consider the following covariates for the initial model:

X_1: "Gender", $X_1 = 1$ for male, and $X_1 = 0$ for female; X_2: "Age", treated as a continuous variable; X_3: "Ethnicity", $X_3 = 1$ if "White, English only", $X_3 = 0$ otherwise; X_4: "Visited doctor since last survey", $X_4 = 1$ if "Yes", $X_4 = 0$ otherwise; X_5: "Describe your health", $X_5 = 1$ if "Very good", $X_5 = 0$ otherwise; X_6: "A measure on depression", $X_6 = 1$ if either "Little interest or pleasure" or "Feeling down or hopeless", $X_6 = 0$ otherwise; X_7: "Frequency of alcohol drinks consumed in the last 12 months", $X_7 = 1$ if "At least one day a week", $X_7 = 0$ otherwise; X_8: "Income categories", $X_8 = 1$ if "Low", $X_8 = 0$ otherwise; X_9: "Education categories", $X_9 = 1$ if "Low", $X_9 = 0$ otherwise; X_{10}: "Marital status", $X_{10} = 1$ if "Married" or "Commonlaw, defacto", $X_{10} = 0$ otherwise; X_{11}: "Mode of data collection", $X_{11} = 1$ if "Internet", $X_{11} = 0$

otherwise.

The data set also contains a column of survey weights for analytical purposes but stratum indicators are not available. In the following analysis, we treat the sample as if it was selected by a single-stage unequal probability sampling design.

We first considered a linear regression model with the estimating functions $g(X_i, Y_i, \theta) = X_i\{Y_i - X_i^{\mathrm{T}}\theta\}$ where $X_i = (1, X_{i1}, \cdots, X_{i11})^{\mathrm{T}}$ and θ is the 12×1 vector of model parameters. We also considered quantile regression models $Q_Y(\gamma \mid X) = X^{\mathrm{T}}\theta_\gamma$ for $\gamma = 0.25$, 0.50 and 0.75 to capture a more complete picture of the effects of the covariates X on the daily cigarette consumption Y. The corresponding estimating functions are $g(X_i, Y_i, \theta_\gamma) = X_i\{\mathrm{I}(Y_i < X_i^{\mathrm{T}}\theta_\gamma) - \gamma\}$. Without loss of generality, we use θ_N to denote the finite population parameters for either the linear regression model or the quantile regression model. The maximum sample empirical likelihood estimator $\hat{\theta}_{SEL}$ of θ_N is presented in Table 2, where the sub-header "Linear Reg." indicates the linear regression model and the other three sub-headers with $\gamma = 0.25$, 0.50 and 0.75 represent the quantile regression models. Also included in the table are the p-values (pval) of the sample empirical likelihood ratio test for H_0: $\theta_{N[j]} = 0$ versus H_1: $\theta_{N[j]} \neq 0$ for each of the 12 components of θ_N, and the tests are done one at a time for $j = 1, 2, \cdots, 12$.

Results in Table 2 provide a preliminary picture on which factors might be important for the model. For the linear regression model, the least significant factor is X_6, "A measure on depression". This seems to be counter-intuitive

to common beliefs. The quantile regression models show different pictures for different γ, and X_6 is indeed significant for $\gamma = 0.25$ at the level of 0.05. Since the tests are done one at a time, a final model cannot be selected from Table 2 unless one uses an iterative method such as stepwise variable selection procedures.

We further consider variable selection using the penalized sample empirical likelihood method. The tuning parameter τ_n is chosen by minimizing the proposed BIC(τ_n) through a fine grid search. The maximum penalized sample empirical likelihood estimates of θ_N for the linear model and the three quantile regression models are presented in Table 3. The final selected models with non-zero coefficients are slightly different for the four models but they all involve a much smaller set of covariates. None of the covariates X_2 (Age), X_7 (Alcohol drinks) and X_{11} (Mode of data collection) is selected in any of the models and X_9 (Education categories) is included for all final models. Other significant factors include X_3 (Ethnic background), which is a bit of surprise, and X_8 (Income categories). The finding that alcohol drinking is unrelated to the heaviness of smoking is also a surprise since drinking and smoking are often believed to go hand-by-hand.

7. Additional Remarks

Survey data are one of the main sources of information for official statistics where descriptive finite population parameters are of primary interest and design-based inferences have been the foundation for survey data analysis. How-

ever, there have been increased use of complex surveys for analytical studies involving statistical models, especially for researchers in social sciences and health and medical fields. The estimating equations approach was first championed by Binder (1983) and Godambe and Thompson (1986). It provides an unified framework for both descriptive and analytical use of survey data and has become a standard tool for both survey researchers and survey data users.

Theoretical developments on survey weighted estimating equations focus mostly on point estimators and variance estimation, and the involved estimating functions are differentiable with the same dimension as the parameters. Binder and Patak (1994) provided a result on confidence intervals for a scalar parameter in the presence of nuisance parameters. In the existing survey sampling literature, the general case over-identified estimating equations system with non-differentiable estimating functions and on general linear or nonlinear hypothesis tests cannot be found. However, there are results which are not as general. Berger and Torres (2016) considered non-differentiable estimating functions for a scalar parameter. This result has been extended for multidimensional parameter by Oguz-Alper and Berger (2016) when the estimating functions are differentiable. Over-identified estimating equations are also considered by Berger and Torres (2016) and Oguz-Alper and Berger (2016) but they are specified by calibration constraints and are not as general as the over-identified system considered here. Wang and Opsomer (2011) discussed variance estimation for parameters involv-

ing non-differentiable estimating functions but the work focused primarily on a scalar parameter.

Variable selection techniques have been extensively discussed in several areas of statistics for model-based inferences. While the same issue of building a model with a large number of covariates is faced by the use of complex survey data, the topic has not been formally discussed under general settings for design-based inferences. Wang, Wang and Wang (2014) is among the first to discuss variable selection for longitudinal survey data using a penalized survey-weighted GEE method.

The sample empirical likelihood methods for complex surveys and the design-based oracle variable selection theory are a general statistical tool for analysis of complex survey data. Zhao and Wu (2018) extended the results presented in this paper to the pseudo empirical likelihood (Chen and Sitter, 1999; Wu and Rao, 2006), and compared the performances of the sample empirical likelihood with the pseudo empirical likelihood through simulation studies. Our proposed design-based variable selection method using the penalized sample empirical likelihood is particularly appealing. It takes into account the sampling design features through the survey weighted estimating equations, which ensures design-consistency for point estimation, and carries over for variable selection through the penalty terms. Design-based variance estimation is not required for variable selection. For large scale complex survey data, the original inclu-

sion probabilities π_i are typically not available. Instead, final adjusted and/or calibrated survey weights are included as part of the public-use survey data. Extending our proposed sample empirical likelihood methods to this practically important topic is currently under investigation.

Supplementary Material

The online Supplementary Material contains technical details and proofs to major theoretical results presented in the main paper and additional simulation results.

Acknowledgement

This research was supported by grants from the Natural Sciences and Engineering Research Council (NSERC) of Canada and a Collaborative Research Team Grant from the Canadian Statistical Sciences Institute (CANSSI). Zhao's research was also supported by the scientific research fund for high-level talents of Yunnan University and the National Natural Science Foundation of China (Grant Nos.: 11731011, 11871287).

References

Berger, Y. G. and 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.

Berger, Y. G. and De La Riva Torres, O. (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.

Berger, Y. G. and De La Riva Torres, O. (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.

Berger, Y. G. (2016) Empirical Likelihood Inference for the Rao-Hartley-Cochran Sampling Design. *Scandinavian Journal of Statistics.* **43**, 721–735.

Berger, Y. G. (2018) An Empirical Likelihood Approach Under Cluster Sampling with Missing Observations. *Annals of the Institute of Statistical Mathematics.* doi.org/10.1007/s10463-018-0681-x.

Binder, D. A. (1983). On the Variances of Asymptotically Normal Estimators from Complex Surveys. *International Statistical Review.* **51**, 279–292.

Binder, D. A. and Patak, Z. (1994). Use of Estimating Functions for Estimation from Complex Surveys. *Journal of the American Statistical Association.* **89**, 1035–1043.

Binder, D. A. and Roberts, G. (2009). Design- and Model-based Inference for Model Parameters. In *Handbook of Statistics*, Volume 29B, *Sample Surveys: Inference and Analysis*, editors: D. Pfeffermann and C.R. Rao, 33–54.

Bowden, R. J. and Turkington, D. A. (1984). *Instrumental Variables.* Cambridge University Press, Cambridge, U.K..

Chen, X., Linton, O. B. and Van Keilegom, I. (2003). Estimation of Semiparametric Models When the Criterion Function is not Smooth. *Econometrica.* **71**, 1591–1608.

Chen, J. and Qin, J. (1993). Empirical Likelihood Estimation for Finite Populations and the Effective Usage of Auxiliary Information. *Biometrika.* **80**, 107–116.

Chen, J. and 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, S. and Kim, J. K. (2014). Population Empirical Likelihood for Nonparametric Inference in Survey Sampling. *Statistica Sinica.* **24**, 335–355.

Fan, J. and Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *Journal of the American statistical Association.* **96**, 1348–1360.

Fan, J. and Peng, H. (2004). Nonconcave Penalized Likelihood with Diverging Number of Parameters. *The Annals of Statistics.* **32**, 928–961.

Francisco, C. A. and Fuller, W. A. (1991). Quantile Estimation with a Complex Survey Design. *The Annals of Statistics.* **19**, 454–469.

Gelman, A. (2007). Struggles with Survey Weighting and Regression Modeling. *Statistical Science.* **22**, 153–164.

Godambe, V. P. and Thompson, M. E. (1986). Parameters of Superpopulation and Survey Population: Their Relationships and Estimation. *International Statistical Review.* **54**,

127–138.

Hartley, H. O. and Rao, J. N. K. (1968). A New Estimation Theory for Sample Surveys. *Biometrika.* **55**, 547–557.

Haziza, D., Mecatti, F. and Rao, J. N. K. (2008). Evaluation of Some Approximate Variance Estimators under the Rao-Sampford Unequal Probability Sampling Design. *Metron.* **66**, 91–108.

Kim, J. K. (2009). Calibration Estimation Using Empirical Likelihood in Survey Sampling. *Statistica Sinica.* **19**, 145–157.

Oguz-Alper, M. and 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.

Pakes, A. and Pollard, D. (1989). Simulation and the Asymptotics of Optimization Estimators. *Econometrica.* **57**, 1027–1057.

Parente, P. M. and Smith, R. J. (2011). GEL Methods for Nonsmooth Moment Indicators. *Econometric Theory.* **27**, 74–113.

Pfeffermann, D. (1993). The Role of Sampling Weights When Modeling Survey Data. *International Statistical Review.* **61**, 317–337.

Qin, J. and Lawless, J. (1994). Empirical Likelihood and General Estimating Equations. *The Annals of Statistics.* **22**, 300–325.

Qin, J. and Lawless, J. (1995). Estimating Equations, Empirical Likelihood and Constraints on Parameters. *The Canadian Journal of Statistics.* **23**, 145–159.

Rao, J. N. K. and Wu, C. (2010). Pseudo Empirical Likelihood Inference for Multiple Frame Surveys. *Journal of the American Statistical Association.* **105**, 1494–1503.

Tang, C. Y. and Leng, C. (2010). Penalized High Dimensional Empirical Likelihood. *Biometrika.* **97**, 905–920.

Thompson, M. E., Fong, G. T., Hammond, D., et al. (2006). Methods of the International Tobacco Control (ITC) Four Country Survey. *Tobacco Control*, 15 (suppl III), iii12-18.

van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes.* Springer, New York.

Wang, J. C. and Opsomer, J. D. (2011). On Asymptotic Normality and Variance Estimation for Nondifferentiable Survey Estimators. *Biometrika.* **98**, 91–106.

Wang, L., Wang, S. and Wang, G. (2014). Variable Selection and Estimation for Longitudinal Survey Data. *Journal of Multivariate Analysis.* **130**, 409–424.

Wu, C. and Rao, J. N. K. (2006). Pseudo Empirical Likelihood Ratio Confidence Intervals for Complex Surveys. *The Canadian Journal of Statistics.* **34**, 359–375.

Zhao, P. and Wu, C. (2018) Some Theoretical and Practical Aspects of Empirical Likelihood Methods for Complex Surveys. *International Statistical Review*, **87**, 239–256.

P. Zhao, D. Haziza and C. Wu

Yunnan Key Laboratory of Statistical Modelling and Data Analysis, Yunnan University, Kunming 650091, China.

E-mail: pyzhao@live.cn

Department of Mathematics and Statistics, Université de Montréal, Montréal, QC, H3T 1J4, Canada

haziza@dms.umontreal.ca

Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON, N2L 3G1, Canada

cbwu@uwaterloo.ca

(Received March 15, 2018; accepted July 7, 2020)

Table 1: Size and Power of the SEL Ratio Test Under Heteroscedasticity

		$\tau = 0.25$			$\tau = 0.5$			$\tau = 0.75$		
Design	b	$N(0,1)$	$\chi^2(3)$	$t(3)$	$N(0,1)$	$\chi^2(3)$	$t(3)$	$N(0,1)$	$\chi^2(3)$	$t(3)$
I	0.50	0.952	0.774	0.909	0.982	0.502	0.975	0.975	0.316	0.893
	0.75	0.498	0.312	0.402	0.578	0.200	0.504	0.553	0.118	0.400
	1.00	0.070	0.069	0.047	0.052	0.060	0.046	0.060	0.055	0.055
	1.25	0.477	0.264	0.377	0.616	0.167	0.499	0.527	0.103	0.326
	1.50	0.952	0.728	0.898	0.993	0.536	0.966	0.970	0.284	0.841
II	0.50	0.926	0.794	0.847	0.989	0.537	0.981	0.984	0.283	0.932
	0.75	0.417	0.335	0.371	0.632	0.193	0.620	0.602	0.123	0.421
	1.00	0.075	0.064	0.060	0.059	0.056	0.053	0.060	0.057	0.049
	1.25	0.516	0.315	0.399	0.660	0.181	0.537	0.588	0.140	0.400
	1.50	0.971	0.789	0.905	0.994	0.557	0.969	0.988	0.381	0.909
III	0.50	0.951	0.743	0.897	0.977	0.492	0.964	0.978	0.306	0.885
	0.75	0.501	0.305	0.397	0.553	0.180	0.505	0.553	0.126	0.393
	1.00	0.069	0.081	0.051	0.055	0.057	0.056	0.052	0.057	0.048
	1.25	0.472	0.252	0.380	0.638	0.168	0.510	0.504	0.120	0.315
	1.50	0.946	0.729	0.886	0.987	0.530	0.968	0.977	0.286	0.830
IV	0.50	0.978	0.863	0.914	0.989	0.530	0.977	0.977	0.291	0.879
	0.75	0.564	0.339	0.423	0.564	0.186	0.516	0.545	0.126	0.391
	1.00	0.061	0.055	0.055	0.056	0.060	0.050	0.055	0.067	0.057
	1.25	0.557	0.331	0.378	0.653	0.224	0.487	0.526	0.125	0.312
	1.50	0.982	0.865	0.895	0.993	0.600	0.966	0.973	0.307	0.815

Table 2: ITC Data: Point Estimates and Tests for $H_0 : \theta_{N[j]} = 0$ vs $H_1 : \theta_{N[j]} \neq 0$

	Linear Reg.		$\gamma = 0.25$		$\gamma = 0.50$		$\gamma = 0.75$	
X	$\hat{\theta}_{SEL}$	pval	$\hat{\theta}_{SEL}$	pval	$\hat{\theta}_{SEL}$	pval	$\hat{\theta}_{SEL}$	pval
1	10.532	0.000	3.000	0.000	10.948	0.350	17.352	0.476
X_1	1.501	0.036	0.387	0.109	3.309	0.000	2.145	0.004
X_2	0.077	0.005	0.032	0.000	0.087	0.000	0.083	0.000
X_3	3.179	0.016	5.548	0.000	2.889	0.005	1.899	0.003
X_4	0.729	0.399	0.774	0.023	0.403	0.248	2.244	0.007
X_5	-2.221	0.004	-1.452	0.007	-2.399	0.006	-2.728	0.000
X_6	0.135	0.865	-1.065	0.036	-0.287	0.235	0.545	0.354
X_7	-0.920	0.215	-0.678	0.064	-1.486	0.074	-1.857	0.008
X_8	1.186	0.224	1.677	0.043	0.544	0.474	0.125	0.780
X_9	1.775	0.017	1.645	0.008	2.015	0.000	2.032	0.013
X_{10}	-0.931	0.244	0.580	0.077	-2.121	0.005	-1.583	0.000
X_{11}	-1.304	0.062	-0.420	0.433	-1.587	0.018	-2.815	0.000

Table 3: ITC Data: Variable Selection with Penalized SEL

X	Linear Reg.	$\gamma = 0.25$	$\gamma = 0.50$	$\gamma = 0.75$
1	13.856	2.664	10.136	6.173
X_1	0.000	0.000	4.025	12.288
X_2	0.000	0.000	0.000	0.000
X_3	3.145	5.486	4.319	0.000
X_4	0.000	1.743	0.000	0.000
X_5	-2.450	0.000	0.000	0.000
X_6	0.000	0.000	0.000	25.069
X_7	0.000	0.000	0.000	0.000
X_8	2.072	2.301	6.706	0.000
X_9	1.871	2.691	2.488	22.962
X_{10}	0.000	0.000	-3.315	0.000
X_{11}	0.000	0.000	0.000	0.000