

# SAMPLE EMPIRICAL LIKELIHOOD AND THE DESIGN-BASED ORACLE VARIABLE SELECTION THEORY

Puying Zhao, David Haziza and Changbao Wu

*Yunnan University, University of Ottawa and University of Waterloo*

*Abstract:* The sample empirical likelihood approach provides a powerful tool for the analysis of complex survey data. We present sample empirical likelihood results for point estimation and linear or nonlinear hypothesis tests on finite population parameters. These parameters are defined using just-identified or over-identified estimating equation systems with smooth or nondifferentiable 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. We investigate the finite sample performances of the proposed methods for quantile regression and variable selection using simulation studies. Lastly, we apply these methods to a survey data set from the International Tobacco Control (ITC) Policy Evaluation Project 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 data collection tool in many areas of scientific investigation. Survey data are widely used for official statistics, social science research, and population health studies. Design-based inferences are the predominant approach in official statistics for descriptive population parameters such as the population mean and quantiles. Survey data are also increasingly being used for analytical purposes, such as exploring the relations between variables or building statistical models for estimation and prediction. However, the use of survey weights in analytical studies has been a subject of debate over the past three decades. One of the central concepts of valid model-based inferences based on survey data is the ignorability of the survey design features. Pfeffermann

---

Corresponding author: Changbao Wu, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON, N2L 3G1, Canada. E-mail: [cbwu@uwaterloo.ca](mailto:cbwu@uwaterloo.ca).

(1993) and Gelman (2007) provide a stimulating discussion on this topic.

The design-based estimating equations approach has gained in popularity among survey researchers and survey data users. The estimating functions are motivated by inferential problems for the superpopulation model parameters  $\theta$ , and the finite-population parameters  $\theta_N$  are defined as the solution to the so-called census estimating equations. Inferences are carried out using the survey-weighted estimating equations. The survey-weighted estimators  $\hat{\theta}$  are typically design-consistent for the finite-population parameters  $\theta_N$ , regardless of the model, and are also valid estimators for the model parameters  $\theta$  if the model holds for the finite population (Godambe and Thompson (1986)). Design-based variance estimators are also valid for estimating model parameters under the joint randomization of the superpopulation model and the survey sampling design (Binder and Roberts (2009)).

The empirical likelihood was first studied by Owen (1988) for independent data, and has since become one of the fastest growing topics in statistics. Owing to its nonparametric features, the method has been discussed extensively in the survey sampling literature for design-based inferences. The first use of the 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 when they used it to estimate 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, focusing on a point estimation for the population mean. Wu and Rao (2006) proposed pseudo empirical likelihood ratio confidence intervals that are applicable to a single parameter under arbitrary sampling designs. Rao and Wu (2010) then extended the method to multiple-frame surveys.

Chen and Kim (2014) proposed the population empirical likelihood approach for parameters defined using estimating equations. They focused on Poisson sampling and conditional Poisson sampling, and established an optimal property for 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) introduced the sample empirical likelihood (SEL) approach, which is a variation of the population empirical likelihood. The SEL estimator is algebraically equivalent to the nonparametric likelihood estimator introduced by Kim (2009). Berger and De La Riva 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 asymptotically follows a standard chi-square

distribution. They considered settings where the parameters are defined using an estimating equations system that is over-identified with calibration constraints. They illustrate their results for four commonly used unequal probability sampling designs. Nonsmooth estimating equations are considered in Berger and De La Riva 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 (2020) addressed issues with nonresponse under cluster sampling. A key feature of the survey designs considered by Berger and his co-authors is that the design-based variance can be approximated without involving second-order inclusion probabilities. The standard chi-square limiting distribution does not hold for arbitrary sampling designs.

This study provides a unified treatment of the SEL approach to design-based survey data analysis. We consider the most general setting, where the vector of the finite-population parameters is defined using a just-identified or over-identified census estimating equations system with smooth or nondifferential estimating functions, with or without additional calibration constraints. The main theoretical results are established under a general setting with an arbitrary unequal probability sampling design. However, unknown joint-inclusion probabilities may be needed for testing in order to estimate the eigenvalues. Approximating these probabilities is often inevitable (e.g., Haziza, Mecatti and Rao (2008)) and requires assumptions about the design, such as high entropy or negligible sampling fractions, as in Berger and De La Riva Torres (2016) and Oguz-Alper and Berger (2016) (see Section 4 in the Supplement Material for more details). Our study makes four major methodological contributions to design-based survey data analyses. First, we establish the design-consistency and asymptotic normality of the maximum SEL estimator. Second, we develop SEL ratio tests for a general linear or nonlinear hypothesis on finite population parameters. Third, we provide a rigorous treatment on parameters defined using nondifferentiable estimating functions, with the general design-based results from the first two contributions covering advanced inference problems, such as quantile regression (QR) analysis. Fourth, we provide a penalized SEL method for design-based variable selection using complex survey data, and establish its oracle properties for parameters defined using general estimating equations. The SEL formulation also yields a computational unification of design-based inferences for surveys and mainstream empirical likelihood applications (Owen (1988); Qin and Lawless (1994, 1995)). Our asymptotic development uses the theory of empiri-

cal processes, and extends existing methods for nonsmooth 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. General results for the SEL on point estimation and linear or nonlinear hypothesis tests are presented in Section 2. Our proposed penalized SEL method for design-based variable selection and its oracle properties are given in Section 3. Practical implementations of the methods are discussed in Section 4. The finite-sample performance of the methods for QR is examined using simulation studies, and is reported in Section 5. An application using a survey data set from the International Tobacco Control (ITC) Policy Evaluation Project is presented in Section 6 to demonstrate the effectiveness of the variable selection method. Additional remarks are given in Section 7. Proofs of the major results, further technical details, regularity conditions, computational details, and additional simulation results are reported in the Supplementary Material.

## 2. SEL Inference for Complex Surveys

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

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

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 using (2.1). Moreover, statistical inferences for commonly encountered model parameters can be carried out using the finite-population parameters defined in (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 nondifferentiable in  $\theta$ . There are three scenarios for over-identified estimating equations systems. In the first scenario, 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 that of an 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  $\theta$ . 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 De La Riva 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 between specific scenarios, and are developed under the general setting.

Few works examine finite-population parameters defined using nondifferentiable estimating functions in the existing literature on survey data analysis. The work of Berger and De La Riva Torres (2016) is based on nondifferentiable estimations defining a scalar parameter. The Oguz-Alper and Berger (2016) multidimensional test is based on profiling, which can be implemented easily under differentiability, as in Qin and Lawless (1994). In Berger and De La Riva Torres (2016) and Oguz-Alper and Berger (2016), differentiability is not needed for point estimation or for testing the whole parameter  $\theta_N$ . For nondifferentiable  $g(X, Y, \theta)$ , exact solutions to  $U_N(\theta) = 0$  may not exist. Under such scenarios, we may replace (2.1) with  $U_N(\theta) = a_N$ , for some sequence  $a_N = O(N^{-1})$ . The introduction of  $a_N$  is for convenience in asymptotic developments involving nondifferentiable estimating functions, and has no practical implications. An important application involving nondifferentiable estimating functions is that of QR analysis using survey data with  $g(X, Y, \theta) = X\{I(Y < X^T\theta) - \tau\}$ , where  $\tau \in (0, 1)$  and  $I(\cdot)$  is the indicator function. We aim to develop a unified theory that covers both smooth and nondifferentiable estimating functions.

## 2.1. Point estimation and asymptotic properties

Let  $\mathcal{S}$  be the set of sampled units selected using a probability sampling design, with first- and second-order inclusion probabilities  $\pi_i = P(i \in \mathcal{S})$  and  $\pi_{ij} =$

$P(i, j \in \mathcal{S})$ , respectively. Let  $n$  be the realized sample size, which can be random under certain sampling designs, such as Poisson sampling. Let  $n_B = E(n \mid \mathcal{F}_N)$  be the expected sample size. Issues with unit or item nonresponses are not considered in this study. The survey sample data set is denoted by  $\{(X_i, Y_i), i \in \mathcal{S}\}$ .

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

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

Note 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  $\pi_i^{-1}$  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 require design-based variance estimation for a general sampling design. The formulation yields a computational unification between the SEL in survey sampling and empirical likelihood methods in other areas, as well as permitting the advanced asymptotic development presented in this paper. Note that our SEL formulation follows Chen and Kim (2014), but equivalent formulations are also used in Berger and De La Riva Torres (2012, 2014, 2016) and Oguz-Alper and Berger (2016). Chen and Kim (2014) rely on Poisson sampling. The use of  $\pi_i^{-1}$  as part of the constraints also appears in the formulation used by Kim (2009), which is, however, different.

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

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

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

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

Finding the solution  $\lambda = \lambda(\theta)$  to (2.2) is a dual problem of maximizing  $l_n(\theta, \lambda)$  with respect to  $\lambda$  for the given  $\theta$ . The maximum SEL estimator  $\hat{\theta}_{SEL}$  for  $\theta_N$  is

given by

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

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

Our first major theoretical result is on design consistency and the asymptotic normality of the maximum SEL estimator  $\hat{\theta}_{SEL}$ . The required regularity conditions C1–C6 are specified in the Supplementary Material. 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, Linton and Van Keilegom (2003)) for independent data to dependent complex survey data. Let  $\|A\| = \{\operatorname{trace}(A^\top A)\}^{1/2}$  for any matrix or vector  $A$ .

**Theorem 1.** *Suppose that Conditions C1–C6 in the Supplementary Material hold. We have the following:*

- (i) *The maximum SEL estimator  $\hat{\theta}_{SEL}$  is design consistent for  $\theta_N$ ; that is, for any  $\epsilon > 0$ ,*

$$\lim_{N \rightarrow \infty} \mathbb{P} \left\{ \|\hat{\theta}_{SEL} - \theta_N\| > \epsilon \mid \mathcal{F}_N \right\} = 0.$$

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

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

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,  $W = n_B N^{-2} \sum_{i=1}^N \pi_i^{-1} [g_i(\theta_N)][g_i(\theta_N)]^\top$ ,  $\Omega = \operatorname{Var}[\hat{U}_N(\theta_N) \mid \mathcal{F}_N]$ , and  $\hat{U}_N(\theta) = N^{-1} \sum_{i \in \mathcal{S}} \pi_i^{-1} g_i(\theta)$ .

Note that the factor  $n_B N^{-2}$  in the definition of  $W$  is not needed to define  $V_1$  because it all cancels out. The design-based variance  $\Omega$  may depend on joint inclusion probabilities. Applications of Theorem 1 and other general results presented in this section require estimating quantities such as  $\Gamma$ ,  $W$ , and  $\Omega$ , which is 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  $V_2 = \Gamma^{-1} \Omega (\Gamma^\top)^{-1}$ .*

- (ii) *Under single-stage PPS sampling with replacement or single-stage PPS sam-*

pling without replacement with negligible sampling fractions, the variance-covariance matrix  $V_1$  reduces to  $V_3 = (n_B \Gamma^\top W^{-1} \Gamma)^{-1}$ .

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

## 2.2. Empirical likelihood ratio tests for general hypotheses on $\theta_N$

We first present the asymptotic distribution of the SEL ratio statistic for  $\theta_N$  under the general setting and an arbitrary sampling design. The SEL ratio statistic for  $\theta_N$  is defined as

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

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 in the Supplementary Material hold. Then, as  $N \rightarrow \infty$ ,  $T_n(\theta_N)$  converges in distribution to  $Q^\top \Delta Q$  when  $\theta_N$  is the true value of the vector parameter, where  $Q$  follows the standard multivariate normal distribution  $N(0, I_r)$  and*

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

The asymptotic distribution of  $T_n(\theta_N)$  can be alternatively represented by  $\sum_{j=1}^p \delta_j \chi_j^2$ , where  $\chi_j^2$ , for  $j = 1, \dots, p$ , are independent random variables, all following the same distribution as  $\chi^2$  with one degree of freedom, and  $\delta_j$ , for  $j = 1, \dots, p$ , are the nonzero eigenvalues of the  $r \times r$  matrix  $\Delta$ . For the special case  $r = p = 1$ , the parameter  $\theta$  becomes a scalar, and  $\Delta$  reduces to a constant  $a = \text{Var}\{\sum_{i \in \mathcal{S}} \pi_i^{-1} g_i(\theta_N) \mid \mathcal{F}_N\} / \sum_{i=1}^N \pi_i^{-1} [g_i(\theta_N)]^2$ . In general, the test statistic  $T_n(\theta_N)$  follows a scaled  $\chi^2$  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 in the Supplementary Material hold. Under single-stage PPS sampling with replacement or single-stage PPS sampling without replacement with negligible sampling fractions, the SEL ratio statistic  $T_n(\theta_N)$  converges in distribution to a  $\chi^2$  random variable with  $p$  degrees of freedom as  $N \rightarrow \infty$ , where  $p$  is the dimension of  $\theta_N$ .*

Corollary 2 can also be found in Oguz-Alper and Berger (2016) and in Berger and De La Riva 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  $\theta_N$  can be constructed as

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

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

We now consider SEL 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  $\theta$ , either linear or nonlinear. Let  $\Theta^* = \{\theta \mid \theta \in \Theta \text{ and } \Phi(\theta) = 0\}$  be the restricted parameter space under  $H_0$ . The restricted maximum SEL estimator of  $\theta$  under  $H_0$  is defined as  $\hat{\theta}_{SEL}^* = \operatorname{argmin}_{\theta \in \Theta^*} \sup_{\lambda \in \hat{\Lambda}_n(\theta)} l_n(\theta, \lambda)$ . Let  $\hat{\lambda}_{SEL}^* = \operatorname{arg sup}_{\lambda \in \hat{\Lambda}_n(\hat{\theta}_{SEL}^*)} l_n(\hat{\theta}_{SEL}^*, \lambda)$ . The SEL 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 in the Supplementary Material and the null hypothesis  $H_0: \Phi(\theta_N) = 0$  hold. As  $N \rightarrow \infty$ , we have the following:*

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

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

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

- (ii) *The SEL ratio statistic  $T_n(\theta_N \mid H_0)$  converges in distribution to  $Q^T \Delta^* Q$ , where  $Q \sim N(0, I_r)$  and*

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

Note that  $\Delta^*$  has the same structure as  $\Delta$  from Theorem 2, with the central piece  $\Sigma$  in  $\Delta$  replaced by  $\Sigma - P_1^* = \Sigma \Psi^T (\Psi \Sigma \Psi^T)^{-1} \Psi \Sigma$  for  $\Delta^*$ . Under general settings, the distribution of  $Q^T \Delta^* Q$  is a weighted  $\chi^2$  involving eigenvalues of  $\Delta^*$ . If  $r = p$ ,  $\Delta^* = n_B \Omega^{1/2} (\Gamma^T)^{-1} \Psi^T (\Psi \Sigma \Psi^T)^{-1} \Psi \Gamma^{-1} \Omega^{1/2}$ . 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  $\Phi(\theta)$  defines a sub-parameter. The result in Part (ii) is simplified for single-stage PPS sampling designs.

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

A similar simplified result is 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  $\theta_N$  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  $\theta_N = (\theta_{N_1}^\top, \theta_{N_2}^\top)^\top$  be a partition of the parameters. The most commonly used linear hypothesis for model building is  $H_0: \theta_{N_2} = 0$ , 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  $\theta_{N_1}$  is a vector of the parameters of interest, and  $\theta_{N_2}$  is treated as a vector of nuisance parameters. Let  $\theta = (\theta_1^\top, \theta_2^\top)^\top$  correspond to the same partition as  $\theta_N$ , with dimension  $k$  for  $\theta_{N_1}$  and  $\theta_1$ . For single-stage PPS sampling designs, the  $(1 - \alpha)100\%$  confidence region for  $\theta_{N_1}$  can be constructed as

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

where  $\tilde{\theta}(\theta_1) = (\theta_1^\top, \hat{\theta}_2(\theta_1)^\top)^\top$ ,  $\hat{\theta}_2(\theta_1) = \operatorname{argmin}_{\theta_2} l_n(\theta, \lambda)$  for the given  $\theta_1$ , and  $\tilde{\lambda} = \operatorname{arg sup}_{\lambda \in \hat{\Lambda}_n(\tilde{\theta}(\theta_1))} l_n(\tilde{\theta}(\theta_1), \lambda)$ . For general sampling designs, the cutoff point  $\chi_{1-\alpha}^2(k)$  needs to be replaced by the estimated  $1 - \alpha$  quantile from the weighted  $\chi^2$  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, behavioral and health aspects of the units are measured for the

survey data. The initial stage for model building requires identifying and selecting of important factors for several responses on addiction and quitting behaviors. Variable selection for complex survey data is an important topic that has not been fully addressed in the existing literature.

In a nonsurvey context, the basic setting for variable selection is to identify the variables in a regression model with zero coefficients. For finite-population regression coefficients  $\theta_N$ , defined as the solution to the census estimating equations, the components of  $\theta_N$  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  $\theta_N$  can be treated as zero, corresponding to the zero coefficients in the superpopulation model.

We consider the smoothly clipped absolute deviations (SCAD) penalty  $p_{\tau_n}(\cdot)$  proposed by Fan and Li (2001), with a tuning parameter  $\tau_n$  selected using a data-driven method. To estimate  $\theta_N$  and identify its nonzero components, we propose using the following penalized SEL function:

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

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

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

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

Let  $\theta_{N[j]}$  be the  $j$ th component of  $\theta_N$ , and let  $\mathcal{A} = \{j \mid 1 \leq j \leq p \text{ and } \theta_{N[j]} \neq 0\}$  be the index set of nonzero components. For asymptotic development, two conditions are assumed (Fan and Li (2001)) for the penalty function and the tuning parameter  $\tau_n$ ; see Supplementary Material for further detail.

Without loss of generality, we assume that  $\theta_N = (\theta_{N_1}^\top, \theta_{N_2}^\top)^\top$ , where  $\theta_{N_1}$  consists of  $d$  nonzero components and  $\theta_{N_2} = 0$ . Let  $\hat{\theta}_{PSEL} = (\hat{\theta}_{P_1}^\top, \hat{\theta}_{P_2}^\top)^\top$  be the penalized maximum SEL estimator of  $\theta_N$ , which is the minimizer of (3.1). Using the partition of  $\theta_N$ , we decompose the variance-covariance matrices  $V_j$  ( $j = 1, 2, 3$ ) and  $\Sigma$ , 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  $\hat{\theta}_{PSEL} = (\hat{\theta}_{P_1}^T, \hat{\theta}_{P_2}^T)^T$  are summarized in the following theorem and corollary.

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

- (i) *the penalized maximum SEL estimator  $\hat{\theta}_{P_2}$  for the zero components satisfies  $P(\hat{\theta}_{P_2} = 0 \mid \mathcal{F}_N) \rightarrow 1$ ;*
- (ii) *the estimator  $\hat{\theta}_{P_1}$  for the nonzero components is asymptotically normal, with mean  $\theta_{N_1}$  and variance-covariance matrix*

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

Part (i) of Theorem 4 is the oracle property of design-based variable selection using the penalized SEL 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 SEL function, which is the first part of  $l_{\tau_n}(\theta)$  given by (3.1). It ensures the 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  $V_{1A}$  for  $\hat{\theta}_{P_1}$  is simplified for the following two special cases:*

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

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

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

Our proposed penalized SEL also provides a 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 SEL estimator of  $\theta_N$  is asymptotically equivalent

to the “unpenalized” maximum SEL estimator of Section 2 on the restricted parameter space  $\Theta^* = \{\theta \mid \theta \in \Theta \text{ and } \theta_{N_2} = 0\}$ .

The penalized SEL can be further used as a general tool to conduct hypotheses tests on the  $d \times 1$  nonzero components of  $\theta_N$  as subsequent steps to variable selection. Specifically, we consider the problem of testing the linear hypotheses

$$H_0 : B\theta_{N_1} = 0, \quad H_1 : B\theta_{N_1} \neq 0, \quad (3.2)$$

where the known  $q \times d$  matrix  $B$  satisfies  $BB^T = I_q$ , with fixed  $q$ . We assume that  $q < d$ ; that is, the number of constraints in  $H_0: B\theta_{N_1} = 0$  is smaller than the number of parameters in  $\theta_{N_1}$ , which excludes  $H_0: \theta_{N_1} = 0$  from consideration. Under a model-based parametric likelihood framework, Fan and Peng (2004) studied a similar type of hypothesis test when the number of parameters is diverging with the sample size. The problem in (3.2) includes hypotheses for individual and multiple components of  $\theta_{N_1}$  as special cases. The most common hypothesis is:  $H_0 : \theta_{N_{1j}} = 0, H_1 : \theta_{N_{1j}} \neq 0$ , where  $\theta_{N_{1j}}$  denotes the  $j$ th coordinate of  $\theta_{N_1}$ . The penalized SEL ratio function is computed as

$$T_{\tau_n}(\theta_{N_1} \mid H_0) = -2 \left\{ l_{\tau_n}(\hat{\theta}_{PSEL}) - \min_{\theta: B\theta_1=0} l_{\tau_n}(\theta) \right\}, \quad (3.3)$$

where  $\theta = (\theta_1^T, \theta_2^T)^T$  follows the same partition as  $\theta_N = (\theta_{N_1}^T, \theta_{N_2}^T)^T$ .

**Theorem 5.** *Suppose that Conditions C1–C8 in the Supplementary Material hold. Then, under single-stage PPS sampling with replacement or single-stage PPS sampling without replacement with negligible sampling fractions, the  $T_{\tau_n}(\theta_{N_1} \mid H_0)$  converges in distribution to a  $\chi^2$  random variable with  $q$  degrees of freedom.*

As a direct consequence, a  $(1 - \alpha)100\%$  confidence region for  $\beta_N = B\theta_{N_1}$  can be constructed as

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

#### 4. Practical Implementations

The asymptotic variance of  $\hat{\theta}_{SEL}$  and the asymptotic distributions of the SEL ratio statistics presented in Sections 2 and 3 are derived under a general sampling design for smooth and nondifferentiable estimating functions. Practical implementations of the methods require estimating three major components,  $W$ ,  $\Gamma$ , and  $\Omega$ , which leads to estimating quantities such as  $V_1$ ,  $\Delta$ , and  $\Delta^*$ , respectively.

The first quantity  $W = n_B N^{-2} \sum_{i=1}^N \pi_i^{-1} [g_i(\theta_N)] [g_i(\theta_N)]^T$  can be consistently

estimated as  $\hat{W} = n_B N^{-2} \sum_{i \in \mathcal{S}} \pi_i^{-2} [g_i(\hat{\theta}_{SEL})][g_i(\hat{\theta}_{SEL})]^\top$ . As noted in Section 2, the factor  $n_B N^{-2}$  is included for theoretical purposes, and is not required for the 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  $\Gamma$  requires additional effort. We provide details of the QR models used for the simulation studies in Section 5, where  $g(X, Y, \theta) = X\{\mathbf{I}(Y < X^\top \theta) - \gamma\}$  for a prespecified  $\gamma \in (0, 1)$ .

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

$$\begin{aligned} U(\theta) &= E\{g(X, Y, \theta)\} \\ &= E[X\{P(Y < X^\top \theta | X) - \gamma\}] \\ &= E[X\{F(X^\top \theta | X) - \gamma\}]. \end{aligned}$$

This leads to  $\Gamma(\theta) = \partial U(\theta)/\partial \theta = E[f(X^\top \theta | X) X X^\top]$ . Let  $K(\cdot)$  be a kernel function. The quantity  $\Gamma(\theta)$  with the given  $\theta$  can be estimated using the survey-weighted estimator

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

where  $h$  is the bandwidth for the kernel density estimation.

The estimation of the third quantity  $\Omega = \text{Var}[\hat{U}_N(\theta_N) | \mathcal{F}_N]$  amounts to a 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 SEL methods developed in this study. 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  $\Omega$ . We provide details in the Supplementary Material for three other commonly encountered sampling designs in survey practice. Each of these designs is examined further in the simulation studies.

## 5. Simulation Studies

We present the results for several simulation studies on the finite-sample performance of the proposed methods for point estimation, hypothesis tests, and

variable selection. We focus on QR models in which the estimating functions are nondifferentiable. The topic of QR has attracted increased attention in recent years in terms of alternative regression modeling techniques. The results for the point estimation and hypothesis tests are reported in this section. The results for the variable selection for QR models and from another simulation on linear regression models are reported in the Supplementary Material.

### 5.1. Basic settings and sampling designs

We consider design-based inferences where the finite population is generated from a superpopulation and is fixed for repeated simulation samples. We consider 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; and (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 Material.

### 5.2. Point estimation and hypothesis tests

Our first simulation study investigates the design-based performance of the SEL estimator and the SEL ratio tests for QR models. The finite population is generated from the model

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

where  $Q_\varepsilon(\gamma)$  is the  $\gamma$ th quantile of  $\varepsilon$ . The scale factor  $\sigma(Z_1, Z_2)$  allows for the presence of conditional heteroscedasticity. The true values of the model parameters are 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 consider  $\sigma(Z_1, Z_2) = 1$  and  $\sigma(Z_1, Z_2) = 1 + Z_2$  to explore the effect of conditional heteroscedasticity. For the error term  $\varepsilon$ , we consider three scenarios: (i)  $\varepsilon \sim N(0, 1)$ ; (ii)  $\varepsilon \sim \chi^2(3)$ ; and (iii)  $\varepsilon \sim t(3)$ . Our simulation examines three QR models, corresponding to  $\tau = 0.25, 0.50$ , and  $0.75$ , respectively.

Let  $\theta = (\theta_0, \theta_1, \theta_2)^\top$ ,  $X = (1, Z_1, Z_2)^\top$ , and  $X_i = (1, Z_{1i}, Z_{2i})^\top$ , for  $i = 1, \dots, N$ . The finite-population parameters  $\theta_N(\gamma) = (\theta_{N_0}(\gamma), \theta_{N_1}(\gamma), \theta_{N_2}(\gamma))^\top$  under the QR model are defined using the census estimating equations  $\sum_{i=1}^N g(X_i, Y_i, \theta_N(\gamma)) = 0$ , where  $g(X, Y, \theta) = X\{I(Y < X^\top\theta) - \gamma\}$ . Under model (5.1) with the shifted  $Q_\varepsilon(\gamma)$  for the error term, the true values of  $\theta_N(\gamma) = (\theta_{N_0}(\gamma), \theta_{N_1}(\gamma), \theta_{N_2}(\gamma))^\top$  for the finite population are essentially the same as the superpopulation model parameters  $\theta = (0.5, 1, 1)^\top$ . For the hypothesis tests, we test  $H_0: \theta_{N_1}(\gamma) = 1.0$

Table 1. Size and power of the SEL ratio test under heteroscedasticity.

Design	$b$	$\tau = 0.25$			$\tau = 0.5$			$\tau = 0.75$		
		$N(0, 1)$	$\chi^2(3)$	$t(3)$	$N(0, 1)$	$\chi^2(3)$	$t(3)$	$N(0, 1)$	$\chi^2(3)$	$t(3)$
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

versus  $H_1$ :  $\theta_{N_1}(\gamma) = b$  for  $b \in \{0.5, 0.75, 1.0, 1.25, 1.5\}$  to examine the size and the power of the SEL ratio test.

For each simulated sample, we compute the maximum SEL estimator  $\hat{\theta}_{SEL}$ , and estimate the three unknown quantities  $W$ ,  $\Gamma$ , and  $\Omega$  using the methods described in Section 4. We use the Gaussian kernel function with bandwidth  $h = n^{-1/3}$  to estimate  $\Gamma$ . The size and power of the test are reported in Table 1 under 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  $\sigma(Z_1, Z_2) = 1$  are included in the Supplementary Material. The simulation results are based on  $B = 2,000$  repeated simulation samples, and can be summarized as follows.

- (a) Point estimation: The relative biases of the maximum SEL estimators are uniformly small ( $< 3\%$ ) for all scenarios considered, including skewed error

distributions, heteroscedasticity, and different sampling designs. The RMS values are similar 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 most 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  $b \neq 1.00$ ) demonstrates 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 with general observations from other studies that the presence of heteroscedasticity affects the performance of tests for QR models.

## 6. An Application to ITC Survey Data

The ITC Project conducts longitudinal surveys to measure the effectiveness of national-level tobacco control policies in more than 20 countries that signed and ratified the Framework Convention on Tobacco Control (FCTC). The ITC Project first started in four countries: Canada, the United States, Australia, and the United Kingdom. The first-wave ITC Four Country Survey used a stratified sampling design and conducted telephone interviews of over 2,000 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 behavior, warning labels, advertising and promotion, light/mild brand descriptors, taxation and purchase behavior, stop-smoking medications and alternative nicotine products, cessation and quitting behavior, and 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. Owing 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 SEL 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 were selected using a single-stage unequal probability sampling design.

We first considered a linear regression model with the estimating functions  $g(X_i, Y_i, \theta) = X_i\{Y_i - X_i^T\theta\}$ , where  $X_i = (1, X_{i1}, \dots, X_{i11})^T$  and  $\theta$  is a  $12 \times 1$  vector of model parameters. We also considered quantile regression models  $Q_Y(\gamma | X) = X^T\theta_\gamma$  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  $g(X_i, Y_i, \theta_\gamma) = X_i\{I(Y_i < X_i^T\theta_\gamma) - \gamma\}$ . Without loss of generality, we use  $\theta_N$  to denote the finite-population parameters for either the linear regression model or the QR model. The maximum SEL estimator  $\hat{\theta}_{SEL}$  of  $\theta_N$  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 SEL ratio test for  $H_0: \theta_{N[j]} = 0$  versus  $H_1: \theta_{N[j]} \neq 0$  for each of the 12 components of  $\theta_N$ , and the tests are done one at a time for  $j = 1, 2, \dots, 12$ .

The 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 counterintuitive. The QR models show different pictures for different  $\gamma$ , and  $X_6$  is indeed significant for  $\gamma = 0.25$  at the level of 0.05. Because the tests are done one at a time, a final

Table 2. ITC data: point estimates and tests for  $H_0 : \theta_{N[j]} = 0$  vs  $H_1 : \theta_{N[j]} \neq 0$ .

$X$	Linear Reg.		$\gamma = 0.25$		$\gamma = 0.50$		$\gamma = 0.75$	
	$\hat{\theta}_{SEL}$	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

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 SEL method. The tuning parameter  $\tau_n$  is chosen by minimizing the proposed  $BIC(\tau_n)$  using a fine grid search. The maximum penalized SEL estimates of  $\theta_N$  for the linear model and the three QR models are presented in Table 3. The final selected models with nonzero 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) are 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, because drinking and smoking are often believed to be related.

## 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. However, there is an increase in the 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 proposed by

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

Binder (1983) and Godambe and Thompson (1986). It provides a unified framework for the 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 estimating functions involved 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, no studies examine the general case of an over-identified estimating equations system with nondifferentiable estimating functions or general linear or nonlinear hypothesis tests. However, there are results that are not as general. Berger and De La Riva Torres (2016) considered nondifferentiable estimating functions for a scalar parameter. This result has been extended for multidimensional parameters by Oguz-Alper and Berger (2016) when the estimating functions are differentiable. Over-identified estimating equations are also considered by Berger and De La Riva 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 involving nondifferentiable estimating functions, but their work focused primarily on a scalar parameter.

Variable selection techniques have been discussed extensively in several areas of statistics for model-based inferences. While the same issue of building a model with a large number of covariates occurs when using complex survey data, the

topic has not been discussed formally under general settings for design-based inferences. Wang et al. (2014) were among the first to discuss variable selection for longitudinal survey data using a penalized survey-weighted GEE method.

The SEL methods for complex surveys and the design-based oracle variable selection theory are general statistical tools for the analysis of complex survey data. Zhao and Wu (2019) extended these results to the pseudo empirical likelihood (Chen and Sitter (1999); Wu and Rao (2006)), and compared the performance of the SEL with that of the pseudo empirical likelihood using simulation studies. Our proposed design-based variable selection method using the penalized SEL is particularly appealing. It takes into account sampling design features using the survey-weighted estimating equations, which ensures design-consistency for point estimation, and carries over for variable selection using the penalty terms. Design-based variance estimation is not required for variable selection. For large-scale complex survey data, the original inclusion probabilities  $\pi_i$  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 SEL methods to this practically important topic is currently under investigation.

## Supplementary Material

The online Supplementary Material contains technical details and proofs of the major theoretical results presented here, as well as additional simulation results.

## Acknowledgments

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, as well as by the National Natural Science Foundation of China (Grant Nos.: 11731011, 11871287).

## References

- Berger, Y. G. (2016). Empirical likelihood inference for the Rao-Hartley-Cochran sampling design. *Scandinavian Journal of Statistics* **43**, 721–735.
- Berger, Y. G. (2020). An empirical likelihood approach under cluster sampling with missing observations. *Annals of the Institute of Statistical Mathematics* **72**, 91–121.
- 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. *Southampton Statistical*

- Sciences Research Institute (S3RI Methodology Working Papers)*. 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. In *Contribution to Sampling Statistics* (Edited by P. L. Mecatti and M. G. Ranalli Conti). Springer.
- 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 (Statistical Methodology)* **78**, 319–341.
- 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, Sample Surveys: Inference and Analysis* (Edited by D. Pfeffermann and C. R. Rao) **28B**, 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., Boudreau, C., Driezen, P., Hyland, A. et al. (2006). Methods of the International Tobacco Control (ITC) Four Country Survey. *Tobacco Control* **15** (suppl III), iii12–iii18.
- 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. (2019) Some theoretical and practical aspects of empirical likelihood methods for complex surveys. *International Statistical Review* **87**, 239–256.

Puying Zhao

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

E-mail: pyzhao@live.cn

David Haziza

Department of Mathematics and Statistics, University of Ottawa, Ottawa, ON, K1N 6N5, Canada.

E-mail: dhaziza@uottawa.ca

Changbao Wu

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

E-mail: cbwu@uwaterloo.ca

(Received March 2018; accepted July 2020)