Empirical Likelihood Inference with Public-Use Survey Data

J.N.K. Rao and Changbao Wu

Carleton University and University of Waterloo
jrao34@gogers.com *and* cbwu@uwaterloo.ca

There has been considerable theoretical development in recent years on empirical likelihood inference for complex surveys, including notably the work by Wu and Rao (2006), Chen and Kim (2014) and Berger and De La Riva Torres (2016) on confidence intervals for finite population parameters. Existing approaches to empirical likelihood inference under the design-based framework are not practically useful since they require the first order inclusion probabilities of the survey design as well as the calibration variables and their known population means or totals, which are not reported in the public-use data files and hence are not available to survey data users. In this paper we develop empirical likelihood methods for analyzing public-use survey data that contain only the variables of interest and the final adjusted and calibrated survey weights along with final replication weights. Asymptotic distributions of the empirical likelihood ratio statistics are derived for parameters defined through estimating equations. Finite sample performances of the empirical likelihood ratio confidence intervals, with comparisons to methods based on the estimating equation theory, are investigated through simulation studies. The proposed approaches make empirical likelihood a practically useful tool for users of complex survey data.

1. Empirical Likelihood and Estimating Equations for Complex Surveys

Let $\mathcal{U} = \{1, 2, \cdots, N\}$ be the set of the survey population, where N is the population size. Let (y_i, x_i) be the measures of the study variable y and auxiliary variables x for unit i and let $\mathcal{F}_{\rm N} = \{(y_i, x_i), i = 1, \cdots, N\}$. Let $\{(y_i, x_i), i \in \mathcal{S}\}$ be survey sample data set. In this section we assume that $\pi_i = P(i \in \mathcal{S})$ are available.

There are two major types of analysis for complex survey data: estimation of descriptive population quantities such as the population mean or analytical use of survey data for statistical modelling. Under both scenarios, the finite population parameters $\theta_{\rm N}$ of dimension p can be defined as the solution to the census estimating equations

$$U_{\rm N}(\theta) = \sum_{i=1}^{N} g(x_i, y_i, \theta) = 0\,, \tag{1}$$

where $g(x, y, \theta)$ is an estimating function of dimension r $(\geq p)$. Under normal circumstances we have $r = p$ but over-identified scenarios with $r > p$ do arise in practice due to additional calibration constraints or known moment conditions over certain variables.

Standard empirical likelihood inference with independent observations as introduced by Owen (1988) with parameters defined by estimating equations as discussed by Qin and Lawless

(1994) consists of three ingredients:

$$\ell(\boldsymbol{p}) = \sum_{i\in\mathcal{S}} \log(p_i)\,, \tag{2}$$

$$\sum_{i\in\mathcal{S}} p_i = 1\,, \tag{3}$$

$$\sum_{i\in\mathcal{S}} p_i g(x_i, y_i, \theta) = 0\,, \tag{4}$$

where $\ell(\boldsymbol{p})$ given by (2) is the empirical log-likelihood function and $\boldsymbol{p} = (p_1, \cdots, p_n)$ is the probability measure over the n sampled units, the equation (3) is the normalization constraint to ensure that $\boldsymbol{p}$ is a discrete probability measure, and the equations (4) are the constraints induced by the parameters θ. The use of $\log(p_i)$ implicitly requires that $p_i > 0$.

When the sample data set $\{(y_i, x_i), i \in \mathcal{S}\}$ is obtained from a complex survey, naive applications of the standard empirical likelihood method produce invalid results under the design-based framework. There have been three major modified approaches in the survey sampling literature on using the empirical likelihood method for complex survey data, and their relations to the standard ingredients (2), (3) and (4) can be described as follows.

(1) *The pseudo empirical likelihood approach (PEL)*: Chen and Sitter (1999) suggested to replace $\ell(\boldsymbol{p})$ by $\ell_{\text{PEL0}}(\boldsymbol{p}) = \sum_{i\in\mathcal{S}} d_i \log(p_i)$, where $d_i = \pi_i^{-1}$ are the basic design weights, while constraints (3) and (4) remain unchanged. The use of $\ell_{\text{PEL0}}(\boldsymbol{p})$ is motivated by the fact that $\ell_{\text{PEL0}}(\boldsymbol{p})$ is the Horvitz-Thompson estimator for the "conceptual" census empirical likelihood function $\sum_{i=1}^N \log(p_i)$. Wu and Rao (2006) used a modified version $\ell_{\text{PEL0}}(\boldsymbol{p}) = n\sum_{i\in\mathcal{S}} \tilde{d}_i(\mathcal{S}) \log(p_i)$, where $\tilde{d}_i(\mathcal{S}) = d_i / \sum_{j\in\mathcal{S}} d_j$, which facilitates the construction of the pseudo empirical likelihood ratio confidence intervals for population parameters. Rao and Wu (2010a) expended the method for multiple frame surveys and Rao and Wu (2010b) developed Bayesian pseudo empirical likelihood method for survey data analysis. However, all existing results on pseudo empirical likelihood methods focus primarily on inferences for a scalar parameter. General statistical tools involving a vector of parameters are not available.

(2) *The population empirical likelihood approach (POEL)*: Chen and Kim (2014) defined the population empirical log-likelihood function as $\ell_{\text{POEL}} = \sum_{i=1}^N \log(\omega_i)$ with normalization constraint $\sum_{i=1}^N \omega_i = 1$. The survey data and parameters are forced into the "population system" through the constraints $\sum_{i\in\mathcal{S}} \omega_i \pi_i^{-1} = 1$ and $\sum_{i\in\mathcal{S}} \omega_i \{g(x_i, y_i, \theta)\pi_i^{-1}\} = 0$. Chen and Kim (2014) focused on Poisson sampling and rejective sampling, and the method hasn't been developed for general unequal probability sampling designs or general inferential problems for analytical use of survey data.

(3) *The sample empirical likelihood approach (SEL)*: The method was first mentioned very briefly by Chen and Kim (2014) as a remark but detailed exploration was not pursued in their paper. The idea is to use the standard empirical log-likelihood function $\ell_{\text{SEL0}}(\boldsymbol{p}) = \sum_{i\in\mathcal{S}} \log(p_i)$ from (2) and the standard normalization constraint (3). The constraints induced by the parameters are modified as $\sum_{i\in\mathcal{S}} p_i \{g(x_i, y_i, \theta)\pi_i^{-1}\} = 0$.

A related recent development was presented in the two papers by Berger and Torres (2016) and Oguz and Berger (2016). The empirical log-likelihood function used by Berger and Torres

(2016) is given by $l_{(m)} = \sum_{i\in\mathcal{S}} \log(m_i)$, where the m_i satisfy the so-called design constraint $\sum_{i\in\mathcal{S}} m_i\pi_i = n$. The m_i can be interpreted as survey weights, since the design constraint reduces to $\sum_{i\in\mathcal{S}} m_i = N$ under simple random sampling. The constraints for the parameters are specified as $\sum_{i\in\mathcal{S}} m_i g(x_i, y_i, \theta) = 0$. It can be seen that, if we let $p_i = m_i\pi_i n^{-1}$, the formulation used by Berger and Torres (2016) is equivalent to the sample empirical likelihood approach.

None of the existing empirical likelihood methods can be used for statistical analysis with public-use survey data files since the initial inclusion probabilities π_i are not available, and calibration variables along with their known population totals are typically not given to the end users of the data files. On the other hand, the availability of replication weights for public-use data sets provides a unique opportunity to develop empirical likelihood as a general statistical tool for survey data analysis.

2. Empirical Likelihood Inference with Public-Use Survey Data

Consider the following version of a micro survey data file, which is released by the survey agency for public use:

$$\left\{\left(y_i, x_i, w_i, w_i^{(1)}, \ldots, w_i^{(B)}\right),\ i = 1, 2, \ldots, n\right\},$$

where the y_i and x_i are possibly vector-values survey variables included in the data set, the w_i is the final survey weight for unit i after unit nonresponse adjustment and/or calibration weighting. Also included in the data file are B final replication weights $w_i^{(1)}$, ..., $w_i^{(B)}$ associated with unit i. The detailed survey design information such as the original design weights $d_i = 1/\pi_i$ and the known auxiliary population information are assumed to be unavailable to the users of the data file. It is also assumed that the finite population size N is unknown.

The survey weighted estimating equations for the vector of parameters θ_{N} are given by

$$\hat{U}_n(\theta) = \sum_{i\in\mathcal{S}} w_i\, g(x_i, y_i, \theta) = 0 \tag{5}$$

For standard scenarios where $r = p$, i.e., the number of equations is the same as the number of parameters, the survey weighted estimator $\hat{\theta}_{\text{N}}$ for θ_{N} is the solution to (5). Let $g_i(\theta) = g(x_i, y_i, \theta)$ and assume that $g_i(\theta)$ is a smooth function of θ. The approximate design-based variance of $\hat{\theta}_{\text{N}}$ has the well-known sandwich form

$$Var\big(\hat{\theta}_{\text{N}}\big) \doteq \Gamma^{-1} Var\big\{\hat{U}_n(\theta_{\text{N}}\big\}\big(\Gamma^{-1}\big)',$$

where $\Gamma = \Gamma(\theta_{\text{N}})$ and $\Gamma(\theta) = \sum_{i=1}^{N} \partial g_i(\theta)/\partial\theta$.

We consider smooth estimating functions and allow over-identified estimating equations system with $r \geq p$. Practically useful results for the special case of $r = p$ and for a scalar parameter (i.e., $p = 1$) will be spelled out whenever is possible. For asymptotic development, we assume that there is sequence of finite populations and a sequence of survey designs with both the population size N and the sample size n going to infinity; see Isaki and Fuller (1982) for further detail. Note that θ_{N} refers to the true value of the finite population parameters.

Assumption A. The final survey weights $(w_1, w_2, \ldots, w_n)$ and the finite population values $\mathcal{F}_{\text{N}} = \{(y_i, x_i), i = 1, \cdots, N\}$ are such that $\hat{U}_n(\theta_{\text{N}}) = \sum_{i \in \mathcal{S}} w_i g_i(\theta_{\text{N}})$ is asymptotically normally distributed with mean zero and variance-covariance matrix at the order $O(N^2/n)$.

Let $\hat{\eta}^{(b)}(\theta_{\text{N}}) = \sum_{i \in \mathcal{S}} w_i^{(b)} g_i(\theta_{\text{N}})$ be the replicated versions of $\hat{U}_n(\theta_{\text{N}}) = \sum_{i \in \mathcal{S}} w_i g_i(\theta_{\text{N}})$ using the bth set of replication weights $(w_1^{(b)}, w_2^{(b)}, \ldots, w_n^{(b)})$, $b = 1, 2, \ldots, B$, assuming θ_{N} is a known number.

Assumption B. The replication variance estimator

$$v\big\{\hat{U}_n(\theta_{\text{N}})\big\} = \frac{1}{B} \sum_{b=1}^{B} \Big\{\hat{\eta}^{(b)}(\theta_{\text{N}}) - \hat{U}_n(\theta_{\text{N}})\Big\}\Big\{\hat{\eta}^{(b)}(\theta_{\text{N}}) - \hat{U}_n(\theta_{\text{N}})\Big\}' \tag{6}$$

is a design-consistent estimator for the true variance-covariance matrix $Var\big\{\hat{U}_n(\theta_{\text{N}})\big\}$.

2.1 The pseudo empirical likelihood approach

Let $\tilde{w}_i(\mathbf{S}) = w_i / \sum_{k \in \mathbf{S}} w_k$, $i \in \mathbf{S}$ be the normalized final survey weights. Let the pseudo empirical log-likelihood function be defined as

$$l_{\text{PEL}}(\boldsymbol{p}) = n \sum_{i \in \mathbf{S}} \tilde{w}_i(\mathbf{S}) \, \log(p_i) \, .$$

For the special case of equal final survey weights, we have $\tilde{w}_i(\mathbf{S}) = 1/n$ and $l_{\text{PEL}}(\boldsymbol{p}) = \sum_{i \in \mathbf{S}} \log(p_i)$. Maximizing $l_{\text{PEL}}(\boldsymbol{p})$ subject to the normalization constraint (3), i.e., $\sum_{i \in \mathbf{S}} p_i = 1$, gives $\hat{\boldsymbol{p}} = (\hat{p}_1, \ldots, \hat{p}_n)$, where $\hat{p}_i = \tilde{w}_i(\mathbf{S})$. Let $\hat{\boldsymbol{p}}(\theta) = (\hat{p}_1(\theta), \ldots, \hat{p}_n(\theta))$ be the maximizer of $l_{\text{PEL}}(\boldsymbol{p})$ under the normalization constraint (3) and the parameter constraint (4), i.e., $\sum_{i \in \mathbf{S}} p_i \, g_i(\theta) = 0$, for a fixed value of θ. It can be shown that

$$\hat{p}_i(\theta) = \frac{\tilde{w}_i(\mathbf{S})}{1 + \lambda g_i(\theta)} \, ,$$

where the Lagrange multiplier λ is the solution to

$$g_{\text{PEL}}(\lambda) = \sum_{i \in \mathbf{S}} \frac{\tilde{w}_i(\mathbf{S}) g_i(\theta)}{1 + \lambda g_i(\theta)} = 0 \, , \tag{7}$$

which can be solved using the modified Newton-Raphson method presented in Chen, Sitter and Wu (2002) and the R code described in Wu (2005). The maximum pseudo empirical likelihood estimator $\hat{\theta}$ is the maximizer of

$$l_{\text{PEL}}\big\{\hat{\boldsymbol{p}}(\theta)\big\} = n \sum_{i \in \mathbf{S}} \tilde{w}_i(\mathbf{S}) \log\big\{\hat{p}_i(\theta)\big\}$$

with respect to θ, which is the same as the solution to

$$\sum_{i \in \mathbf{S}} \hat{p}_i \, g_i(\theta) = \sum_{i \in \mathbf{S}} \tilde{w}_i(\mathbf{S}) \, g_i(\theta) = 0 \, ,$$

since it achieves the global maximum under the normalization constraint. For the population mean $\theta_N = \mu_y = N^{-1} \sum_{i=1}^{N} y_i$, the maximum pseudo empirical likelihood estimator is given by $\hat{\theta} = \sum_{i \in \mathbf{S}} \tilde{w}_i(\mathbf{S}) y_i = \sum_{i \in \mathbf{S}} w_i y_i / \sum_{i \in \mathbf{S}} w_i$.

Our primary interest is to construct pseudo empirical likelihood ratio confidence intervals for θ with the public-use survey data file. The pseudo empirical log-likelihood ratio statistic for θ is given by

$$r_{\text{PEL}}(\theta) = l_{\text{WR}}\big\{\hat{\boldsymbol{p}}(\theta)\big\} - l_{\text{WR}}\big(\hat{\boldsymbol{p}}\big) = -n\sum_{i\in\mathbf{S}} \tilde{w}_i(\mathbf{S}) \log\big\{1+\lambda g_i(\theta)\big\}\,.$$

Under the regularity conditions described in Wu and Rao (2006) on the final survey weights w_i and the variable $u_i = g_i(\theta_N)$, we can show that $\lambda = O\big(n^{-1/2}\big)$, $\max_{i\in\mathbf{S}} |\lambda g_i(\theta_N)| = o_p(1)$ and

$$\lambda = \Big\{\sum_{i\in\mathbf{S}} \tilde{w}_i(\mathbf{S}) g_i(\theta)\Big\} \Big/ \Big[\sum_{i\in\mathbf{S}} \tilde{w}_i(\mathbf{S})\big\{g_i(\theta)\big\}^2\Big] + o_p\Big(n^{-1/2}\Big)\,.$$

This leads to the following asymptotic expansion to the pseudo empirical log-likelihood ratio statistic:

$$-2r_{\text{PEL}}(\theta) = n\Big\{\sum_{i\in\mathbf{S}} \tilde{w}_i(\mathbf{S}) g_i(\theta)\Big\}^2 \Big/ \Big[\sum_{i\in\mathbf{S}} \tilde{w}_i(\mathbf{S})\Big\{g_i(\theta)\Big\}^2\Big] + o_p\big(1\big)\,,$$

which further leads to the following major asymptotic result. Note that the adjusting factor $\hat{a}_{\text{PEL}}$ used in the theorem is of order $O_p(1)$.

Theorem 1. *Under Assumptions 1 and 2, the adjusted pseudo empirical log-likelihood ratio statistic* $-2r_{PEL}(\theta)/\hat{a}_{PEL}$ *converges in distribution to a* χ^2 *random variable with one degree of freedom when* $\theta = \theta_N$*, where the adjusting factor* $\hat{a}_{PEL}$ *is computed as*

$$\hat{a}_{PEL} = v\big\{\hat{U}_n(\hat{\theta})\big\} \Big/ \Big[n^{-1}\hat{N}\sum_{i\in\mathbf{S}} w_i\big\{g_i(\hat{\theta})\big\}^2\Big]\,,$$

with $v\big\{\hat{U}_n(\hat{\theta})\big\}$ *being the replication variance estimator given in Assumption 2 but replacing* θ_N *by* $\hat{\theta}$*, and* $\hat{N} = \sum_{i\in\mathbf{S}} w_i$*.*

It is important to notice that the adjusting factor $\hat{a}_{\text{PEL}}$ as well as $-2r_{\text{PEL}}(\theta)$ for a given θ can be computed based solely on the public-use survey data file. No additional information is required. The $1-\alpha$ level pseudo empirical likelihood ratio confidence interval for θ_N can therefore be constructed as

$$\mathcal{C}_1 = \Big\{\ \theta\ \Big|\ -2r_{\text{PEL}}(\theta)/\hat{a}_{\text{PEL}} \le \chi^2_1(\alpha)\Big\}\,, \tag{8}$$

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

2.2 The sample empirical likelihood approach

The sample empirical likelihood approach can be adapted for public-use survey data. We start with the standard empirical log-likelihood function

$$l_{\text{SEL}}(\boldsymbol{p}) = \sum_{i\in\mathbf{S}} \log(p_i)\,.$$

Maximizing $l_{\text{SEL}}(\boldsymbol{p})$ under the normalization constraint (3), i.e., $\sum_{i\in\mathbf{S}} p_i = 1$, gives $\hat{p}_i = n^{-1}$, $i \in \mathbf{S}$. The constraint for the parameter θ defined through (1) is formed using the transformed variable $w_i g_i(\theta)$ and is given by

$$\sum_{i\in\mathbf{S}} p_i\big\{w_i g_i(\theta)\big\} = 0\,. \tag{9}$$

Let $\hat{\boldsymbol{p}}(\theta) = (\hat{p}_1(\theta), \ldots, \hat{p}_n(\theta))$ be the maximizer of $l_{\text{SEL}}(\boldsymbol{p})$ under the normalization constraint (3) and the parameter constraint (9) for a fixed θ. It follows from standard empirical likelihood method that

$$\hat{p}_i(\theta) = \frac{1}{n}\,\frac{1}{1+\lambda\{w_i g_i(\theta)\}}$$

for $i \in \mathbf{S}$, where the Lagrange multiplier λ is the solution to the equation

$$g_{\text{SEL}}(\lambda) = \frac{1}{n}\sum_{i\in\mathbf{S}} \frac{w_i g_i(\theta)}{1+\lambda\{w_i g_i(\theta)\}} = 0\,.$$

The empirical log-likelihood ratio statistic for θ under the current setting is given by

$$r_{\text{SEL}}(\theta) = l_{\text{BT}}\big\{\hat{\boldsymbol{p}}(\theta)\big\} - l_{\text{BT}}\big(\hat{\boldsymbol{p}}\big) = \sum_{i\in\mathbf{S}} \log\{n\hat{p}_i(\theta)\} = -\sum_{i\in\mathbf{S}} \log\{1+\lambda w_i g_i(\theta)\}\,.$$

It can be shown that

$$-2r_{\text{SEL}}(\theta) = \Big\{\sum_{i\in\mathbf{S}} w_i g_i(\theta)\Big\}^2 \Big/ \Big[\sum_{i\in\mathbf{S}}\big\{w_i g_i(\theta)\big\}^2\Big] + o_p(1)\,.$$

Theorem 2. *Under Assumptions 1 and 2, the adjusted empirical log-likelihood ratio statistic* $-2r_{BT}(\theta)/\hat{a}_{BT}$ *converges in distribution to a* χ^2 *random variable with one degree of freedom when* $\theta = \theta_N$*, where the adjusting factor* $\hat{a}_{BT}$ *is computed as*

$$\hat{a}_{BT} = v\big\{\hat{U}_n(\hat{\theta})\big\} \Big/ \Big[\sum_{i\in\mathbf{S}}\big\{w_i g_i(\hat{\theta})\big\}^2\Big]\,,$$

with $v\big\{\hat{U}_n(\hat{\theta})\big\}$ *being the replication variance estimator given in Assumption 2 but replacing* θ_N *by* $\hat{\theta}$*.*

It should be emphasized once again that the adjusting factor $\hat{a}_{\text{SEL}}$ and the empirical likelihood ratio function $-2r_{\text{SEL}}(\theta)$ for a given θ can be computed based on the public-use survey data file. No additional information is required. The $1-\alpha$ level empirical likelihood ratio confidence interval for θ_N can be constructed as

$$\mathcal{C}_2 = \Big\{\ \theta\ \Big|\ -2r_{\text{SEL}}(\theta)/\hat{a}_{\text{SEL}} \le \chi^2_1(\alpha)\Big\}\,. \tag{10}$$

REFERENCES: Omitted!