Statistica Sinica **14**(2004), 1057-1067

SOME ALGORITHMIC ASPECTS OF THE EMPIRICAL LIKELIHOOD METHOD IN SURVEY SAMPLING

Changbao Wu

University of Waterloo

Abstract: Recent development of the empirical likelihood method in survey sampling has attracted attention from survey statisticians. Practical considerations for using the method in real surveys depend largely on the availability of simple and efficient algorithms for computing the related weights for the maximum empirical likelihood estimators. In this article we briefly describe the modified Newton-Raphson procedure of Chen, Sitter and Wu (2002) for non-stratified sampling designs and show that under suitable reformulation the algorithm can also be used to handle stratified sampling, the most commonly used design in survey practice. The idea of the q-weighted empirical likelihood approach is briefly introduced and the related algorithms are discussed. The proposed algorithms are tested in a limited simulation study and are shown to perform well.

Key words and phrases: Newton-Raphson algorithm, pseudo empirical likelihood, stratified sampling, weighted empirical likelihood.

1. Introduction

Since Owen's pioneering work (1988, 1990), the empirical likelihood (EL) method has emerged in the past decade as a powerful inference tool with promising applications in many areas of statistics. Historically, however, the first application of the concept behind empirical likelihood was suggested by Hartley and Rao (1968) in survey sampling. Under simple random sampling, they developed the so-called scale-load estimator based on the multinomial likelihood function and showed that known auxiliary information can be incorporated through constrained maximum likelihood estimation. The first formal application of the EL method in survey sampling, following the work of Owen, was presented by Chen and Qin (1993) where the log-likelihood function under independent observations, i.e., $l_{1.1}(\boldsymbol{p}) = \sum_{i\in s}\log(p_i)$, was used for simple random sampling with or without replacement. Here $\boldsymbol{p}$ represents $\{p_i,\, i \in s\}$. Zhong and Rao (1996) considered stratified random sampling and used the log-likelihood function $l_{1.2}(\boldsymbol{p}) = \sum_h \sum_{i\in s_h}\log(p_{hi})$, where s_h is the set of sampled units from stratum h, and $\sum_{i\in s_h} p_{hi} = 1,\ h = 1,\ldots,H$. In this case $\boldsymbol{p}$ denotes $\{p_{hi},\, i \in s_h,\ h = 1,\ldots,H\}$. Since the stratum samples s_h are independent, this is a true

empirical log-likelihood function if the s_h are drawn by simple random sampling with replacement.

Generalization of the EL method under unequal probability sampling designs was proposed by Chen and Sitter (1999) using a pseudo empirical likelihood approach. For non-stratified sampling design, they recommended use of the so-called pseudo empirical (log) likelihood function $l_{2.1}(\boldsymbol{p}) = \sum_{i\in s} d_i \log(p_i)$, where $d_i = 1/\pi_i$ are the basic design weights, and $\pi_i = P(i \in s)$ are the inclusion probabilities. They justified the use of $l_{2.1}(\boldsymbol{p})$ as a likelihood function by noting that under the probability sampling design $E\{l_{2.1}(\boldsymbol{p})\} = \sum_{i=1}^{N} \log(p_i)$, the log-likelihood function one would use at the population level if one views the entire finite population as an independent sample from a superpopulation. The pseudo empirical likelihood function $l_{2.1}(\boldsymbol{p})$ can be viewed as a design-based estimate for that total likelihood function. Also note that, under simple random sampling where $d_i = N/n$, $l_{2.1}(\boldsymbol{p})$ differs from $l_{1.1}(\boldsymbol{p})$ only by a multiplying constant.

The pseudo empirical likelihood function $l_{2.1}(\boldsymbol{p})$ can also be motivated through the well-known Kullback-Leibler divergence widely used in information theory. Let $d_i^* = d_i/\sum_{i\in s} d_i$. The KL divergence measure between $\boldsymbol{d}^* = \{d_i^*, i \in s\}$ and $\boldsymbol{p} = \{p_i, i \in s\}$ is defined as $D_{KL}(\boldsymbol{d}^*, \boldsymbol{p}) = \sum_{i\in s} d_i^* \log(d_i^*) - \sum_{i\in s} d_i^* \log(p_i)$. Maximizing $l_{2.1}(\boldsymbol{p})$ is equivalent to minimizing $D_{KL}(\boldsymbol{d}^*, \boldsymbol{p})$ with respect to $\boldsymbol{p}$, a criterion often used for model selection.

Under stratified sampling with arbitrary sampling design within each stratum, the pseudo empirical likelihood function is defined as $l_{2.2}(\boldsymbol{p}) = \sum_h \sum_{i\in s_h} d_{hi} \log(p_{hi})$, where $d_{hi} = 1/\pi_{hi}$, and the π_{hi} are the stratum inclusion probabilities. Note that $l_{2.2}(\boldsymbol{p})$ is a design-based estimate for the total likelihood $\sum_{h=1}^{H} \sum_{i=1}^{N_h} \log(p_{hi})$, where N_h are the stratum sizes. It is interesting to observe that under stratified random sampling where $d_{hi} = N_h/n_h$, $l_{2.2}(\boldsymbol{p})$ does not reduce to $l_{1.2}(\boldsymbol{p})$ unless the stratum sample sizes n_h are allocated proportional to the stratum sizes N_h.

The empirical likelihood method in survey sampling is primarily focused on the use of auxiliary information to construct point estimators for various finite population quantities. Estimation of general parameters defined through estimating equations can be handled as well. Empirical likelihood ratio confidence intervals are usually difficult to use under complex sampling designs due to the lack of limiting distributions. See Zhong and Rao (2000) for an example under stratified random sampling. In all cases the method involves constrained maximization of the empirical likelihood function.

One of the major objectives of this article is to present simple and efficient algorithms for computing the weights for the maximum pseudo empirical likelihood estimators under a general probability sampling design. Section 2 provides a short review of the modified Newton-Raphson algorithm of Chen, Sitter and Wu

(2002) for non-stratified sampling. It is shown in Section 3 that this algorithm can be adapted to handle stratified sampling through suitable reformulation of the problem. The idea of the q-weighted empirical likelihood approach is briefly introduced and the related algorithms are discussed in Section 4. The proposed algorithms are tested in Section 5 through a limited simulation study. We conclude with some remarks in Section 6.

2. The Modified Newton-Raphson Algorithm for Non-Stratified Sam pling

Let s be the set of sampled units under a non-stratified sampling design with first order inclusion probabilities π_i, and $d_i = 1/\pi_i$ be the basic design weights. Let $\bar{\boldsymbol{X}} = N^{-1}\sum_{i=1}^{N}\boldsymbol{x}_i$ be the known vector-valued population means of auxiliary variables $\boldsymbol{x}$. The maximum pseudo empirical likelihood estimator for the population mean $\bar{Y} = N^{-1}\sum_{i=1}^{N} y_i$ of the study variable y is defined as $\hat{\bar{Y}}_{PEL} = \sum_{i\in s}\hat{p}_i y_i$, where the weights $\hat{p}_i$ maximize the pseudo empirical log-likelihood function $l_{2.1}(\boldsymbol{p}) = \sum_{i\in s} d_i \log(p_i)$ subject to

$$\sum_{i\in s} p_i = 1 \;\; (p_i > 0) \;\text{ and }\; \sum_{i\in s} p_i\boldsymbol{x}_i = \bar{\boldsymbol{X}}\,. \tag{2.1}$$

The second part of constraint (2.1), $\sum_{i\in s} p_i\boldsymbol{x}_i = \bar{\boldsymbol{X}}$, is often referred to as a benchmark constraint. By using the usual Lagrange multiplier method it can be shown that the solution to the above constrained maximization problem is given by $\hat{p}_i = d_i^*/(1+\boldsymbol{\lambda}'\boldsymbol{u}_i)$, where $d_i^* = d_i/\sum_{i\in s} d_i$, $\boldsymbol{u}_i = \boldsymbol{x}_i - \bar{\boldsymbol{X}}$, and the Lagrange multiplier $\boldsymbol{\lambda}$ is the solution to

$$g(\boldsymbol{\lambda}) = \sum_{i\in s}\frac{d_i^*\boldsymbol{u}_i}{1+\boldsymbol{\lambda}'\boldsymbol{u}_i} = \boldsymbol{0}\,. \tag{2.2}$$

It can be shown that, with probability tending to one as the sample size goes to infinity, there exists a unique solution to (2.2). For a given sample s, the set of feasible values of $\boldsymbol{\lambda}$ such that $\hat{p}_i > 0$ is given by $A = \{\boldsymbol{\lambda} : 1+\boldsymbol{\lambda}'\boldsymbol{u}_i > 0,\ i\in s\}$ which is a convex set. Another key observation leading to the proposed algorithm is that maximizing $l_{2.1}(\boldsymbol{p})$ subject to (2.1) is a dual problem of maximizing the concave function $\tilde{l}(\boldsymbol{\lambda}) = \sum_{i\in s} d_i^*\log(1+\boldsymbol{\lambda}'\boldsymbol{u}_i)$ with respect to $\boldsymbol{\lambda}$ over the convex set A, since $\partial\tilde{l}(\boldsymbol{\lambda})/\partial\boldsymbol{\lambda} = g(\boldsymbol{\lambda})$. If the unique solution to (2.2) exists, it can be found through the following modified Newton-Raphson algorithm of Chen, Sitter and Wu (2002).

Step 0: Let $\boldsymbol{\lambda}_0 = \boldsymbol{0}$. Set $k = 0$, $\gamma_0 = 1$ and $\epsilon = 10^{-8}$.

Step 1: Calculate $\Delta(\boldsymbol{\lambda}_k)$ where

$$\Delta(\boldsymbol{\lambda}) = \left\{\frac{\partial}{\partial \boldsymbol{\lambda}} g(\boldsymbol{\lambda})\right\}^{-1} g(\boldsymbol{\lambda}) = \left\{-\sum_{i\in s} \frac{d_i^* \boldsymbol{u}_i \boldsymbol{u}_i'}{(1+\boldsymbol{\lambda}'\boldsymbol{u}_i)^2}\right\}^{-1} \sum_{i\in s} \frac{d_i^* \boldsymbol{u}_i}{1+\boldsymbol{\lambda}'\boldsymbol{u}_i}\,.$$

If $\|\Delta(\boldsymbol{\lambda}_k)\| < \epsilon$, stop the algorithm and report $\boldsymbol{\lambda}_k$; otherwise go to Step 2.

Step 2: Calculate $\boldsymbol{\delta}_k = \gamma_k \Delta(\boldsymbol{\lambda}_k)$. If $1 + (\boldsymbol{\lambda}_k - \boldsymbol{\delta}_k)'\boldsymbol{u}_i \le 0$ for some i or $\tilde{l}(\boldsymbol{\lambda}_k - \boldsymbol{\delta}_k) < \tilde{l}(\boldsymbol{\lambda}_k)$, let $\gamma_k = \gamma_k/2$ and repeat Step 2.

Step 3: Set $\boldsymbol{\lambda}_{k+1} = \boldsymbol{\lambda}_k - \boldsymbol{\delta}_k$, $k = k+1$ and $\gamma_{k+1} = (k+1)^{-1/2}$. Go to Step 1.

The modification step 2 ensures that at each iteration the value of $\boldsymbol{\lambda}$ is still inside the feasible range A and that the concave function $\tilde{l}(\boldsymbol{\lambda})$ is moving towards its maximum point. The algorithm is simple and efficient, and convergence is guaranteed. It can easily be programmed using any popular statistical software such as SAS or R/Splus.

3. Algorithms for Stratified Sampling

Let y_{hi} and $\boldsymbol{x}_{hi}$ be the values of the y and the $\boldsymbol{x}$ variables for the ith unit in stratum h, respectively, with $h = 1, \ldots, H$ and $i = 1, \ldots, N_h$. Let $W_h = N_h/N$ be the stratum weights where $N = \sum_{h=1}^{H} N_h$ is the population size. Let d_{hi} be the basic design weights for stratum h. The maximum pseudo EL estimator of $\bar{Y} = N^{-1}\sum_h \sum_{i=1}^{N_h} y_{hi}$ is defined as $\hat{\bar{Y}}_{PEL} = \sum_{h=1}^{H} W_h \sum_{i\in s_h} \hat{p}_{hi} y_{hi}$, where the weights $\hat{p}_{hi}$ maximize $l_{2.2}(\boldsymbol{p}) = \sum_{h=1}^{H}\sum_{i\in s_h} d_{hi}\log(p_{hi})$ subject to constraints

$$\sum_{i\in s_h} p_{hi} = 1 \;\; (p_{hi} > 0)\,, \quad h = 1, \cdots, H\,, \tag{3.1}$$

$$\sum_{h=1}^{H} W_h \sum_{i\in s_h} p_{hi}\boldsymbol{x}_{hi} = \bar{\boldsymbol{X}}\,. \tag{3.2}$$

The sub-normalization of weights within each stratum (i.e., constraints (3.1)) is practically important since the strata means $\bar{Y}_h = N_h^{-1}\sum_{i=1}^{N_h} y_{hi}$ are often themselves of interest and can be estimated by $\hat{\bar{Y}}_h = \sum_{i\in s_h} \hat{p}_{hi} y_{hi}$. This can also be justified by noting that the s_h are independent samples from each of the strata. The benchmark constraints (3.2) are used due to two possible reasons: the surveyor is only interested in benchmarking at the population level and/or the stratum means $\bar{\boldsymbol{X}}_h = N_h^{-1}\sum_{i=1}^{N_h} \boldsymbol{x}_{hi}$ are not available. If $\bar{\boldsymbol{X}}_h$ are known and the benchmark constraints are imposed at the stratum level, the problem then reduces to separate maximum EL estimation within each of the strata.

Maximizing $l_{2.2}(\boldsymbol{p})$ subject to (3.1) and (3.2) is not a trivial task. The algorithm of Chen, Sitter and Wu (2002) is not directly applicable under stratified

sampling. While the objective function $l_{2.2}(\boldsymbol{p})$ and the benchmarking constraints (3.2) are both at the population level (double summations), the normalization constraints (3.1) are imposed at the stratum level (single summation). Two possible strategies can be used: (i) set an arbitrary benchmark constraint for each stratum and find the intermediate solution under that constraint, and then seek for the final solution through profile likelihood method, see Section 3.1; (ii) try to reformulate the constraints to put everything on the population level so that the algorithm of Section 2 can directly be applied, see Section 3.2.

3.1. An existing algorithm for stratified sampling

Chen and Sitter (1999) presented an algorithm for computing the $\hat{p}_{hi}$ through the profile likelihood method. This algorithm was also introduced by Zhong and Rao (2000). To start, let $\boldsymbol{\theta}_h$, $h = 1, \ldots, H$ be a group of vectors such that $\sum_{h=1}^{H} W_h \boldsymbol{\theta}_h = \bar{\boldsymbol{X}}$. We first maximize $l_{2.2}(\boldsymbol{p})$ subject to restrictions $\sum_{i \in s_h} p_{hi} = 1$ and $\sum_{i \in s_h} p_{hi} \boldsymbol{x}_{hi} = \boldsymbol{\theta}_h$ for fixed $\boldsymbol{\theta}_h$, $h = 1, \ldots, H$. This amounts to finding maximum EL estimators for each of the strata. The algorithm of Section 2 can now be used. The solution is given by $p_{hi} = d_{hi}/\{d_h + \boldsymbol{\lambda}_h'(\boldsymbol{x}_{hi} - \boldsymbol{\theta}_h)\}$, with $d_h = \sum_{i \in s_h} d_{hi}$, and $\boldsymbol{\lambda}_h$ satisfying

$$\sum_{i \in s_h} \frac{d_{hi}(\boldsymbol{x}_{hi} - \boldsymbol{\theta}_h)}{d_h + \boldsymbol{\lambda}_h'(\boldsymbol{x}_{hi} - \boldsymbol{\theta}_h)} = \boldsymbol{0}\,, \quad h = 1, \ldots, H\,. \tag{3.3}$$

By omitting a constant, the resulting likelihood function for the given set of $\boldsymbol{\theta}_h$ is given by

$$l(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_H) = -\sum_{h=1}^{H} \sum_{i \in s_h} d_{hi} \log[d_h + \boldsymbol{\lambda}_h'(\boldsymbol{x}_{hi} - \boldsymbol{\theta}_h)]\,.$$

We further maximize $l(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_H)$ with respect to $\boldsymbol{\theta}_h$ under the restriction $\sum_{h=1}^{H} W_h \boldsymbol{\theta}_h = \bar{\boldsymbol{X}}$. It can be shown that the final choice of $\boldsymbol{\theta}_h$ satisfies

$$\sum_{i \in s_h} \frac{d_{hi}(x_{hi} - \boldsymbol{\theta}_h)}{d_h + W_h \boldsymbol{t}'(x_{hi} - \boldsymbol{\theta}_h)} = \boldsymbol{0}\,, \quad h = 1, \ldots, H \tag{3.4}$$

for some vector-valued $\boldsymbol{t}$. Theoretically one can search for the final solution through profile analysis: for each possible value of $\boldsymbol{t}$, solve (3.4) to get $\boldsymbol{\theta}_h$, and determine the set of $\boldsymbol{\theta}_h$ that maximize $l(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_H)$, then use the final $\boldsymbol{\theta}_h$ to compute $\hat{p}_{hi}$.

This algorithm is most efficient when the $\boldsymbol{x}$ variable is univariate, since in this case one can search for the final solution by increasing or decreasing the value of $\boldsymbol{t}$, which is also one dimensional, to find the maximum point of $l(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_H)$.

When the $\boldsymbol{x}$ variables are high dimensional, however, computing $\hat{p}_{hi}$ through the profile likelihood method involves repeatedly solving the high dimensional nonlinear systems (3.3) and (3.4), and is not practically achievable. An alternative approach is required.

3.2. The adapted algorithm

The real computational difference between stratified sampling and non-stratified sampling comes from the first set of restrictions, i.e., $\sum_{i\in s} p_i = 1$ versus (3.1). Note that we can rewrite (3.1) in an equivalent form as

$$\sum_{h=1}^{H} W_h \sum_{i\in s_h} p_{hi} = 1\,, \tag{3.5}$$

$$\sum_{i\in s_h} p_{hi} = 1\,,\ h = 1,\ldots,H-1\,. \tag{3.6}$$

If we keep (3.5) separate and combine (3.6) together with (3.2), then the structural difference in terms of constraints between the two classes of sampling designs disappears (it is just a matter of single or double summation!). This can be achieved by augmenting the $\boldsymbol{x}$ variable to include the first $H-1$ strata indicator variables. More specifically, suppose $\boldsymbol{x}_{hi} = (x_{hi1},\ldots,x_{hik})'$ is of dimension k, let

$$\begin{aligned}
\boldsymbol{z}_{1i} &= (1,0,\ldots,0,x_{1i1},\ldots,x_{1ik})'\,,\\
\boldsymbol{z}_{2i} &= (0,1,\ldots,0,x_{2i1},\ldots,x_{2ik})'\,,\\
&\vdots\\
\boldsymbol{z}_{(H-1)i} &= (0,0,\ldots,1,x_{(H-1)i1},\ldots,x_{(H-1)ik})'\,,\\
\boldsymbol{z}_{Hi} &= (0,0,\ldots,0,x_{Hi1},\ldots,x_{Hik})'
\end{aligned}$$

and $\bar{\boldsymbol{Z}} = (W_1,\ldots,W_{H-1},\bar{X}_1,\ldots,\bar{X}_k)'$, where $(\bar{X}_1,\ldots,\bar{X}_k)' = \bar{\boldsymbol{X}}$. The two sets of constraints (3.6) and (3.2) can be combined into a single set of constraints

$$\sum_{h=1}^{H} W_h \sum_{i\in s_h} p_{hi}\boldsymbol{z}_{hi} = \bar{\boldsymbol{Z}}\,. \tag{3.7}$$

Maximizing $l_{2.2}(\boldsymbol{p})$ under the restrictions (3.1) and (3.2) is equivalent to maximizing $l_{2.2}(\boldsymbol{p})$ subject to (3.5) and (3.7). By using the Lagrange multiplier method we can show that the solution is $\hat{p}_{hi} = d^*_{hi}/(1+\boldsymbol{\lambda}'\boldsymbol{u}_{hi})$, where $d^*_{hi} = d_{hi}/\{W_h\sum_{h=1}^{H}\sum_{i\in s_h} d_{hi}\}$, $\boldsymbol{u}_{hi} = \boldsymbol{z}_{hi} - \bar{\boldsymbol{Z}}$, and the Lagrange multiplier $\boldsymbol{\lambda}$ is the solution to

$$\sum_{h=1}^{H}\sum_{i\in s_h} \frac{d_{hi}\boldsymbol{u}_{hi}}{1+\boldsymbol{\lambda}'\boldsymbol{u}_{hi}} = \boldsymbol{0}\,. \tag{3.8}$$

It is immediately evident that (3.8) is computationally identical to (2.2). The algorithm of Chen, Sitter and Wu (2002) described in Section 2 can directly be used for solving (3.8) and is guaranteed to converge to the unique solution if such a solution exists. The key to the above reformulation is to use only $H-1$ indicator variables and keep (3.5) separate, otherwise the problem will not reduce to the non-stratified case.

Practical issues may arise when the total number of strata, H, is too large, since the augmented variable $\boldsymbol{z}$ contains $H-1$ indicator variables. The effect of large H depends on the capacity of the software and hardware systems used in handling high dimensional matrix manipulations. The indicator variables alone will not create ill-conditioned matrices during the updating process of the algorithm. See the comments at the end of Sections 4 and 5 for more discussion on this issue. When H is very large but $\boldsymbol{x}$ is of low dimension, one may choose to use the algorithm of Chen and Sitter (1999). The adapted algorithm, however, is generally applicable. It is simple and yet efficient, and requires solving a non-linear system only once using the well-developed algorithm of Chen, Sitter and Wu (2002). The proposed reformulation brings a unified approach to computing maximum empirical likelihood estimators under both stratified and non-stratified sampling designs.

4. Algorithms for the Q-weighted Empirical Likelihood Method

It has been shown by Chen and Sitter (1999) that the maximum pseudo empirical likelihood estimator $\hat{\bar{Y}}_{PEL}$ is asymptotically equivalent to the calibration estimator $\hat{\bar{Y}}_C = N^{-1}\sum_{i\in s} w_i y_i$, where the calibrated weights w_i minimize the chi-squared distance measure $\Phi_s = \sum_{i\in s}(d_i - w_i)^2/d_i$ between the w_i and the basic design weights d_i subject to $\sum_{i\in s} w_i \boldsymbol{x}_i = \boldsymbol{X}$, with $\boldsymbol{X} = \sum_{i=1}^{N} \boldsymbol{x}_i$ being the known population totals.

Deville and Särndal (1992) presented a general class of calibration estimators and demonstrated that calibration estimators obtained by using different distance measures are asymptotically equivalent to those from using a chi-squared distance measure $\Phi_s^* = \sum_{i\in s}(d_i - w_i)^2/(d_i q_i)$ with certain choice of the q-weights in defining Φ_s^*. Obviously the EL estimator corresponds to $q_i \equiv 1$. An interesting question arises from this context: can a calibration estimator with an arbitrary distance measure (or arbitrary q-weights) be achieved through the empirical likelihood approach? The q-weighted empirical likelihood estimator defined in the sequel does exactly that. Some theoretical properties of this approach have been investigated by Wu (2003). We focus here on the computational aspect of the method.

Let $d_i^* = d_i/\sum_{i\in s} d_i$ be normalized design weights for the given sample under a non-stratified sampling scheme. The q-weighted empirical likelihood function

is defined as $l^*_{2.1}(\boldsymbol{p}) = \sum_{i\in s} q_i^{-1}\{d_i^* \log(p_i) - p_i\}$. The weighted maximum empirical likelihood estimator of $\bar{Y}$ is computed as $\hat{\bar{Y}}_{WEL} = \sum_{i\in s}\hat{p}_i y_i$, where the $\hat{p}_i$ maximize $l^*_{2.1}(\boldsymbol{p})$ subject to the same set of constraints (2.1). This weighted EL approach reduces to the usual EL method under the uniform weights $q_i = 1$, since in this case $l^*_{2.1}(\boldsymbol{p}) = l_{2.1}(\boldsymbol{p}) - 1$ under the constraint $\sum_{i\in s} p_i = 1$.

There are two algorithmic aspects involved here which are different from those of Section 2. First, the Lagrange multiplier corresponding to $\sum_{i\in s} p_i = 1$ cannot be eliminated under the current context. We need to combine $\sum_{i\in s} p_i = 1$ and $\sum_{i\in s} p_i \boldsymbol{x}_i = \bar{\boldsymbol{X}}$. This can be done by augmenting the $\boldsymbol{x}$ variables to include 1 as the first component. Secondly, the centered variables $\boldsymbol{u}_i = \boldsymbol{x}_i - \boldsymbol{X}$ cannot be used in the combined constraints, and the convex duality discussed in Section 2 is no longer the case, as shown below.

Let $\boldsymbol{x}^*_i = (1, \boldsymbol{x}_i')'$ and $\bar{\boldsymbol{X}}^* = (1, \bar{\boldsymbol{X}}')'$. The set of constraints (2.1) is equivalent to $\sum_{i\in s} p_i \boldsymbol{x}^*_i = \bar{\boldsymbol{X}}^*$. Using a straightforward Lagrange multiplier argument it can be shown that $\hat{p}_i = d^*_i/(1 + \boldsymbol{\lambda}'\boldsymbol{x}^*_i q_i)$ for $i \in s$, where the Lagrange multiplier $\boldsymbol{\lambda}$ is the solution to

$$g^*(\boldsymbol{\lambda}) = \sum_{i\in s} \frac{d^*_i \boldsymbol{x}^*_i}{1 + \boldsymbol{\lambda}'\boldsymbol{x}^*_i q_i} - \bar{\boldsymbol{X}}^* = \boldsymbol{0}\,. \tag{4.1}$$

It is the term $-\bar{\boldsymbol{X}}^*$ in $g^*(\boldsymbol{\lambda})$ that destroys the convex duality discussed in the modified Newton-Raphson algorithm. The algorithm, however, can still be used to solve (4.1) under minor modifications. At Step 1, one needs to compute

$$\Delta(\boldsymbol{\lambda}) = \left\{\frac{\partial}{\partial \boldsymbol{\lambda}} g^*(\boldsymbol{\lambda})\right\}^{-1} g^*(\boldsymbol{\lambda}) = \left\{-\sum_{i\in s} \frac{q_i d^*_i \boldsymbol{x}^*_i (\boldsymbol{x}^*_i)'}{(1+\boldsymbol{\lambda}'\boldsymbol{x}^*_i q_i)^2}\right\}^{-1} \left(\sum_{i\in s} \frac{d^*_i \boldsymbol{x}^*_i}{1+\boldsymbol{\lambda}'\boldsymbol{x}^*_i q_i} - \bar{\boldsymbol{X}}^*\right).$$

At Step 2, one only checks if $1 + (\boldsymbol{\lambda}_k - \boldsymbol{\delta}_k)'\boldsymbol{x}^*_i q_i \le 0$ for some i.

Under stratified sampling, the q-weighted empirical likelihood function is defined as $l^*_{2.2}(\boldsymbol{p}) = \sum_h W_h \sum_{i\in s_h} q_{hi}^{-1}\{d^*_{hi}\log(p_{hi}) - p_{hi}\}$, where $d^*_{hi} = d_{hi}/\sum_{i\in s_h} d_{hi}$ and q_{hi} are prespecified q-weights. The use of stratum weights W_h in defining $l^*_{2.2}(\boldsymbol{p})$ is to bring certain consistency between $l_{2.2}(\boldsymbol{p})$ and $l^*_{2.2}(\boldsymbol{p})$: when $q_i = 1$ and $\hat{N}_h = \sum_{i\in s_h} d_{hi} = N_h$, the above defined $l^*_{2.2}(\boldsymbol{p})$ reduces to $l_{2.2}(\boldsymbol{p})$ if one ignores a trivial constant term.

The adapted algorithm described in Section 3.2 can be modified to handle the current situation. Note that the Lagrange multiplier corresponding to (3.5) cannot be eliminated, we now need to augment the $\boldsymbol{x}_{hi}$ variable to include all H stratum indicator variables, denoted by $\boldsymbol{z}_{hi}$. Let $\bar{\boldsymbol{Z}} = (W_1, \ldots, W_H, \bar{\boldsymbol{X}}')'$. It follows that the two sets of constraints (3.1) and (3.2) are equivalent to $\sum_{h=1}^H W_h \sum_{i\in s_h} p_{hi}\boldsymbol{z}_{hi} = \bar{\boldsymbol{Z}}$. The weighted maximum empirical likelihood estimator is computed as $\hat{\bar{Y}}_{WEL} = \sum_{h=1}^H W_h \sum_{i\in s_h} \hat{p}_{hi} y_{hi}$, where $\hat{p}_{hi} = d^*_{hi}/(1+\boldsymbol{\lambda}'\boldsymbol{z}_{hi}q_{hi})$

and the Lagrange multiplier $\boldsymbol{\lambda}$ is the solution to

$$g^*(\boldsymbol{\lambda}) = \sum_{h=1}^{H} \sum_{i \in s_h} \frac{W_h d^*_{hi} \boldsymbol{z}_{hi}}{1 + \boldsymbol{\lambda}' \boldsymbol{z}_{hi} q_{hi}} - \bar{\boldsymbol{Z}} = \boldsymbol{0}\,. \tag{4.2}$$

The solution to (4.2) can be found using an algorithm similar to the one for solving (4.1).

It should be noted that algorithms presented in this section can be used to compute the usual unweighted EL estimators by simply setting $q_i = 1$. It is easy to argue in this context that the total number of strata, H, has little impact on the performance of the adapted algorithm for stratified sampling, as long as the high dimensional matrices involved can be handled by the computing facilities used. The key step of the algorithm involves the inversion of the matrix $M = \sum\sum W_h q_{hi} d^*_{hi} \boldsymbol{z}_{hi} \boldsymbol{z}'_{hi} (1 + \boldsymbol{\lambda}' \boldsymbol{z}_{hi} q_{hi})^{-2}$. The $H \times H$ submatrix in the upper-left corner of M is diagonal and is bounded away from zero.

5. A Simulation

We test our proposed algorithms through a limited simulation study. Six stratified finite populations are generated from the regression model $y_{hi} = \alpha_h + \beta_h x_{hi} + \gamma_h x^a_{hi} \varepsilon_{hi}$, $i = 1, \ldots, N_h$, $h = 1, 2, 3, 4$, where $\alpha_h = 2h$, $\beta_h = h$ and $\gamma_h = rh$. The stratum sizes are chosen as $N_h = 2000 - 400h$ for $h = 1, 2, 3, 4$. The covariates x_{hi} are generated from $\chi^2(2h)$. Two different types of distributions are used to generate the error terms ε_{hi}: the skewed standard log-normal distribution $LN(0, 1)$, and the symmetric standard normal distribution $N(0, 1)$. The six finite populations correspond to the parameter setting $a = 0.0$, 0.5 and 1.0 for each of the two error distributions. The control parameter r used in $\gamma_h = rh$ is chosen such that the finite population correlation coefficient between y and x is 0.80.

For each fixed finite population, a stratified random sample under proportional allocation is drawn where the stratum sample sizes are given by $n_h = 100 - 20h$ for $h = 1, 2, 3, 4$. For each simulated sample, the unweighted estimator (PEL) and the q-weighted estimators (WEL) are computed. The weighting schemes used are $q_{hi} = x^{-2a}_{hi}$. Performance of these estimators is evaluated in terms of Relative Bias (RB) and Relative Efficiency (RE) defined as $RB = B^{-1} \sum_{b=1}^{B} \{\hat{\bar{Y}}(b) - \bar{Y}\}/\bar{Y}$ and $RE = MSE(\hat{\bar{Y}}_0)/MSE(\hat{\bar{Y}})$, where $MSE(\hat{\bar{Y}}) = B^{-1} \sum_{b=1}^{B} \{\hat{\bar{Y}}(b) - \bar{Y}\}^2$, $\hat{\bar{Y}}(b)$ is the estimator $\hat{\bar{Y}}$ under study computed from the bth simulated sample, and $\hat{\bar{Y}}_0 = \sum_h W_h \bar{y}_h$ is the stratified mean estimator used for baseline comparison. The process is independently repeated $B = 1,000$ times.

Table 1 presents the simulated relative efficiencies (RE) for estimators under investigation. The absolute values of relative biases (RB) for all cases are less than 0.2% and are not reported here. Under non-stratified sampling it has been

observed by Wu (2003) that the two estimators PEL and WEL have similar performance under large samples, but the weighted estimator WEL performs much better when the sample size is small or moderate. Given that the overall sample size is 200 here, our results seem to support a similar argument under stratified sampling.

Table 1. Simulated Relative Efficiency (RE).

	$\varepsilon \sim LN(0,1)$		$\varepsilon \sim N(0,1)$	
a	PEL	WEL	PEL	WEL
0.0	2.06	2.06	1.79	1.79
1.0	1.92	1.94	1.85	1.86
2.0	1.87	1.80	1.80	1.76

The simulation study was conducted using algorithms described in Sections 3 and 4 and programmed in R/Splus. The R codes are available from the author. Non-convergence never occurred for the unweighted EL estimators. A few cases of singular matrices occurred during the process of updating the gradient $\Delta(\boldsymbol{\lambda}_k)$ for the q-weighted EL estimator, and the related samples were dropped from the simulation.

We also examined the proposed algorithms under another scenario where only a few elements are sampled from each stratum but the total number of strata is large. We generated stratified finite populations using a similar regression model as before, but in this case each stratum consisted of ten clusters, with each cluster having four elements. A stratified single stage cluster sampling scheme was used to select two psu (clusters) per stratum. For a population with H strata, the population size is $40H$ and the overall sample size is $8H$. Using a dual process Sun Unix workstation with 768 megabytes of memory, the adapted algorithm programmed in R can handle H as large as 400. In most cases the algorithm converges within six iterations regardless of the value of H.

6. Concluding Remarks

The most commonly used method for the analysis of large scale surveys by many statistical agencies is the generalized regression estimator or, equivalently, the calibration estimator under the chi-squared distance measure. It is an effective way of achieving the goal of benchmarking and yet is relatively simple in terms of computation. The weights produced through the generalized regression method, however, can take negative values. This drawback has long been recognized. Ad hoc adjustments to this problem have been proposed but none of them is universally accepted. The empirical likelihood method is an attractive

alternative approach. In addition to its clear maximum likelihood interpretation, it can also be viewed as a calibration approach with desirable features such as high efficiency and intrinsic positive weights. The computational algorithms presented in this article make the method easily implementable in practice and, perhaps, more popular among survey statisticians.

Acknowledgements

This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada. The author gratefully acknowledges the constructive comments and suggestions of the editor, an associate editor and a referee.

References

Chen, J. and Qin, J. (1993). Empirical likelihood estimation for finite population 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. *Statist. Sinica* **9**, 385-406.

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

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

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

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

Owen, A. B. (1990). Empirical likelihood confidence regions. *Ann. Statist.* **18**, 90-120.

Wu, C. (2003). Weighted empirical likelihood inference. Working paper 2003-01, Department of Statistics and Actuarial Science, University of Waterloo.

Zhong, C. X. Bob and Rao, J. N. K. (1996). Empirical likelihood inference under stratified random sampling using auxiliary information. *Proceedings of the Section on Survey Research Methods*, ASA, 798-803.

Zhong, B. and Rao, J. N. K. (2000). Empirical likelihood inference under stratified random sampling using auxiliary population information. *Biometrika* **87**, 929-938.

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

E-mail: cbwu@uwaterloo.ca

(Received February 2003; accepted June 2003)