Catalogue no. 11-522-XIE

Statistics Canada International Symposium Series - Proceedings

Symposium 2003: Challenges in Survey Taking for the Next Decade

2003

Statistics Canada Statistique Canada

Canada

Proceedings of Statistics Canada Symposium 2003
Challenges in Survey Taking for the Next Decade

COMBINING INFORMATION FROM MULTIPLE SURVEYS THROUGH THE EMPIRICAL LIKELIHOOD METHOD

Changbao Wu[1]

ABSTRACT

It is often desirable to combine information collected in compatible multiple surveys so that certain consistency requirements are met and more efficient estimators are obtained. Earlier work on this topic includes Zieschang (1990) and Renssen and Nieuwenbroek (1997) who suggested to use the generalized regression estimator with enlarged number of auxiliary variables to achieve that goal. A drawback of their approach is that the adjusted weights can take negative values which is very undesirable. In this article the author uses the recently developed pseudo empirical likelihood method to construct estimators which not only meet the consistency and efficiency requirements but have more attractive features. The two approaches are asymptotically equivalent but the latter has clear maximum likelihood interpretations and the resulting adjusted weights are always positive. He also provides efficient algorithms for computing the proposed estimators and thus makes the method easily applicable for real surveys.

KEYWORDS: Benchmark Constraints; Consistency Requirements; Generalized Regression Estimator; Newton-Raphson Algorithm; Pseudo Empirical Likelihood

1. INTRODUCTION

In survey practice weight adjustment is routinely performed to accommodate, among other things, internal consistency requirements that are of interest to both the survey statisticians and the potential users of the survey data. Benchmark constraints are most commonly imposed where the adjusted weights w_i reproduce the known population totals (or means) of auxiliary variables $\mathbf{x}$, i.e. $\sum_{i\in s} w_i \mathbf{x}_i = \mathbf{X}$, where s represents the set of sampled units and $\mathbf{X}$ is the vector of known population totals. Such an adjustment can be achieved by using the generalized regression estimator. The generalized regression estimator is not only a vehicle to achieve the benchmark constraints but also more efficient when compared to the baseline Horvitz-Thompson estimator.

When two (or more) surveys are conducted for the same target population, another consistency requirement may arise. If some auxiliary variables are jointly collected in both surveys but their population totals are unknown, then it is desirable that, in addition to benchmark constraints over auxiliary variables with known population totals, the weights of both surveys produce the same estimates for the unknown population totals of the common auxiliary variables. This problem has previously been addressed by Zieschang (1990) and Renssen and Nieuwenbroek (1997). They both proposed to use the generalized regression (GREG) estimator with enlarged number of auxiliary variables to achieve that goal.

The generalized regression approach, however, has an undesirable property which was already being recognized by the authors: "A disadvantage of the method is the increased possibility of negative weights, due to the enlarged number of explanatory variables. The occurrence of negative weights is inherent to the general regression estimator, and for many users this is an undesirable feature." (Renssen and Nieuwenbroek, 1997).

We propose to use the recently developed pseudo empirical likelihood (EL) method to construct estimators that not only meet the efficiency and consistency requirements but have more attractive features. The EL method is a powerful nonparametric inference tool with applications in many areas of statistics. See Owen (2001) for a comprehensive account and updated overview of the subject. Historically, however, this method was first used in

[1] Changbao Wu, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada, N2L 3G1, cbwu@uwaterloo.ca

survey sampling by Hartley and Rao (1968). Its discrete and nonparametric nature is particularly appealing for finite population problems. In this article we demonstrate that the EL approach is well-suited under the current context, and consistency and efficiency requirements between two or multiple surveys can naturally be formed as constraints and be integrated into the maximum likelihood estimation process. The two approaches, GREG and EL, are asymptotically equivalent but the latter has clear maximum likelihood interpretations and the resulting weights are always positive.

We consider two surveys in what follows but our method can be extended to handle multiple surveys. A logically sound approach involves a joint maximum likelihood estimation using two samples. This is presented in Section 2. Also in Section 2, we present two algorithms for computing the proposed EL estimator. The first algorithm involves profile likelihood method in searching for a solution and is efficient only when the common auxiliary variable is of dimension one. The second algorithm employs a novel reformulation of the problem and can easily be applied under general situations using the well developed algorithm of Chen, Sitter and Wu (2002). In Section 3, a separate empirical likelihood approach is employed where the EL estimators are computed separately for each survey with the unknown population means of the common auxiliary variables estimated from the combined sample data and used as control values. Computation in this case is simple and straightforward. The finite sample performance of the proposed EL estimators, with comparison to the GREG estimators of Zieschang (1990) and of Renssen and Nieuwenbroek (1997), is investigated in Section 4 through a simulation study. We conclude some remarks in Section 5.

2. THE COMBINED EL APPROACH

Suppose the finite population consists of N identifiable units. Associated with the ith unit are values of the study variables y_1 and y_2, and the vectors of auxiliary variables $\mathbf{x}_1, \mathbf{x}_2$ and $\mathbf{z}$, denoted by $y_{1i}, y_{2i}, \mathbf{x}_{1i}, \mathbf{x}_{2i}$, and $\mathbf{z}_i$, respectively, for i = 1,2,…,N. Information on $(y_1, \mathbf{x}_1)$ is collected in the first survey and information on $(y_2, \mathbf{x}_2)$ is gathered in the second survey. In addition, data on the common auxiliary variables $\mathbf{z}$ are collected in both surveys. The two surveys, however, are carried out independent of each other. The two sets of sample data are $\{(y_{1i}, \mathbf{x}_{1i}, \mathbf{z}_i), i \in s_1\}$ and $\{(y_{2j}, \mathbf{x}_{2j}, \mathbf{z}_j), j \in s_2\}$, where s_1 and s_2 are the sets of sampled units from the first and the second survey, respectively. The population means $\overline{\mathbf{X}}_t = N^{-1}\sum_{i=1}^{N}\mathbf{x}_{ti}$ are known (t = 1, 2) but $\overline{\mathbf{Z}} = N^{-1}\sum_{i=1}^{N}\mathbf{z}_i$ are unknown. Zieschang (1990) and Renssen and Nieuwenbroek (1997) provided excellent motivations and real examples on this setting including a highly valuable application on the split questionnaire survey designs. Let $\overline{Y}_t = N^{-1}\sum_{i=1}^{N} y_{ti}, t = 1, 2,$ be the population quantities of interest. If y_1 and y_2 measure the same characteristic but over different time periods, then the difference $\Delta = \overline{Y}_2 - \overline{Y}_1$ may also be of interest.

Following similar arguments as in Chen and Sitter (1999), the combined pseudo empirical log-likelihood function based on the two samples can be written as

$$l(\mathbf{p}, \mathbf{q}) = \sum_{i \in s_1} d_{1i} \log(p_i) + \sum_{j \in s_2} d_{2j} \log(q_j),$$

where $\mathbf{p} = (p_1, ..., p_{n_1})', \mathbf{q} = (q_1, ..., q_{n_2})', p_i = \Pr(y_1 = y_{1i}), q_j = \Pr(y_2 = y_{2j}), d_{ti} = 1/\pi_{ti}, \pi_{ti}$ are the first order inclusion probabilities and n_t is the sample size from the tth survey, t = 1, 2.

The maximum pseudo empirical likelihood estimators for $\overline{Y}_1$ and $\overline{Y}_2$ are defined as

$$\hat{\overline{Y}}_1 = \sum_{i \in s_1} \hat{p}_i y_{1i} \text{ and } \hat{\overline{Y}}_2 = \sum_{j \in s_2} \hat{q}_j y_{2j},$$

where the $\hat{p}_i$ and $\hat{q}_j$, which are interpreted as the adjusted weights, maximize the joint pseudo empirical likelihood function $l(\mathbf{p}, \mathbf{q})$ subject to a system of normalization and consistency requirements:

$$\sum_{i \in s_1} p_i = 1(p_i > 0), \sum_{j \in s_2} q_j = 1(q_j > 0), \tag{1}$$

$$\sum_{i \in s_1} p_i \mathbf{x}_{1i} = \overline{\mathbf{X}}_1, \sum_{j \in s_2} q_j \mathbf{x}_{2j} = \overline{\mathbf{X}}_2, \tag{2}$$

$$\sum_{i \in s_1} p_i \mathbf{z}_i = \sum_{j \in s_2} q_j \mathbf{z}_j, \tag{3}$$

The two sets of benchmark constraints in (2) could involve measurements on the same **x** variables and hence the same population means as well. In the absence of known population means, some or all of the equations in (2) can be removed from the system. The last set of equations (3) brings consistency between the two surveys over the common auxiliary variables. They also make the resulting estimators $\hat{\overline{Y}}_1$ and $\hat{\overline{Y}}_2$ more efficient by using the combined information from both surveys.

One of the related issues here is the existence of the foregoing defined combined EL estimators. The maximum pseudo empirical likelihood estimators $\hat{\overline{Y}}_t$ will not exist if $\overline{\mathbf{X}}_t$ is not an inner point of the convex hull formed by $\{\mathbf{x}_{ti}, i \in s_t\}$, or if the two convex hulls formed by $\{\mathbf{z}_i, i \in s_1\}$ and $\{\mathbf{z}_j, j \in s_2\}$ are disjoint. This occurs with probability approaching to zero as both sample sizes go to infinity. A proof of this can be sketched by following the lines of Lemma 1 of Chen and Sitter (1999).

Another practically important issue is the computational aspect of the proposed EL method. We present two algorithms, both of them take advantage of the well-behaved algorithm of Chen *et al*. (2002) for computing maximum empirical likelihood estimators under a single nonstratified sample. The first algorithm is efficient when the common auxiliary variable is univariate, while the second algorithm can be used under general situations.

2.1 The first algorithm

Let $\sum_{i \in s_1} p_i \mathbf{z}_i = \sum_{j \in s_2} q_j \mathbf{z}_j = \boldsymbol{\theta}$ be fixed. It is then straightforward to show by using the Lagrange multiplier method that

$$\hat{p}_i = \frac{d_{1i}^*}{1 + \boldsymbol{\lambda}_1' \mathbf{u}_{1i}(\boldsymbol{\theta})}, \quad \hat{q}_j = \frac{d_{2j}^*}{1 + \boldsymbol{\lambda}_2' \mathbf{u}_{2j}(\boldsymbol{\theta})}, \tag{4}$$

where $d_{ti}^* = d_{ti} / \sum_{i \in s_t} d_{ti}$ and

$$\mathbf{u}_{ti}(\boldsymbol{\theta}) = \begin{pmatrix} \mathbf{x}_{ti} - \overline{\mathbf{X}}_t \\ \mathbf{z}_{ti} - \boldsymbol{\theta} \end{pmatrix},$$

with the understanding that $\mathbf{z}_{ti}$ refers to $\mathbf{z}_i$ from the *t*th sample, *t* = 1, 2. The Lagrange multipliers $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$ are the solutions to

$$\sum_{i \in s_1} \frac{d_{1i}^* \mathbf{u}_{1i}(\boldsymbol{\theta})}{1 + \boldsymbol{\lambda}_1' \mathbf{u}_{1i}(\boldsymbol{\theta})} = 0 \text{ and } \sum_{j \in s_2} \frac{d_{2j}^* \mathbf{u}_{2j}(\boldsymbol{\theta})}{1 + \boldsymbol{\lambda}_2' \mathbf{u}_{2j}(\boldsymbol{\theta})} = 0, \tag{5}$$

respectively. The profile likelihood function for $\boldsymbol{\theta}$ is then obtained by putting $\hat{p}_i$ and $\hat{q}_j$ into $l(\mathbf{p},\mathbf{q})$ and is given by (with a constant term omitted)

$$l(\boldsymbol{\theta}) = -\sum_{i\in s_1} d_{1i}\,\log\{1+\boldsymbol{\lambda}_1'\,\mathbf{u}_{1i}(\boldsymbol{\theta})\} - \sum_{j\in s_2} d_{2j}\,\log\{1+\boldsymbol{\lambda}_2'\,\mathbf{u}_{2j}(\boldsymbol{\theta})\}.$$

The maximum point of $l(\boldsymbol{\theta})$, denoted by $\hat{\boldsymbol{\theta}}$, can be found through the conventional profile analysis. The final adjusted weights $\hat{p}_i$ and $\hat{q}_j$ are obtained by plugging $\hat{\boldsymbol{\theta}}$ and the associated $\boldsymbol{\lambda}_t$ into (4).

This algorithm involves finding $\boldsymbol{\lambda}_t\,(t=1,2)$ as solutions to (5) for each fixed value of $\boldsymbol{\theta}$, and then finding $\hat{\boldsymbol{\theta}}$ that maximizes $l(\boldsymbol{\theta})$. For the first part, a simple and stable algorithm for solving (5) to obtain the vector-valued $\boldsymbol{\lambda}_t$ has been deployed by Chen *et al.* (2002). As for $\hat{\boldsymbol{\theta}}$, if the common auxiliary variable $\mathbf{z}$ is of dimension one, it can easily be found through the usual profile likelihood method. When $\mathbf{z}$ is high dimensional, so is $\boldsymbol{\theta}$, this algorithm becomes awkward and impracticable. A more flexible algorithm is needed.

2.2 The second algorithm

Suppose $\mathbf{z}_i = (z_{1i},\ldots,z_{ki})'$ is of dimension *k*. If we augment $\mathbf{z}_i$ to *k*+1 dimensional by including $z_{(k+1)i}=1$ as the last component, we can rewrite the system of constraints (1), (2) and (3) as

$$\sum_{i\in s_1} p_i + \sum_{j\in s_2} q_j = 2, \tag{6}$$

$$\begin{bmatrix} \mathbf{X}^{(1)} & \mathbf{0} \\ \mathbf{0} & \mathbf{X}^{(2)} \\ \mathbf{Z}^{(1)} & -\mathbf{Z}^{(2)} \end{bmatrix} \begin{bmatrix} \mathbf{p} \\ \mathbf{q} \end{bmatrix} = \begin{bmatrix} \overline{\mathbf{X}}_1 \\ \overline{\mathbf{X}}_2 \\ \mathbf{0} \end{bmatrix}, \tag{7}$$

where $\mathbf{X}^{(t)} = (\mathbf{x}_{t1},\ldots,\mathbf{x}_{tn_t}), \mathbf{Z}^{(t)} = (\mathbf{z}_{t1},\ldots,\mathbf{z}_{tn_t}), \mathbf{z}_{ti}$ represents $\mathbf{z}_i$ from the *t*th survey with 1 as its last component, *t* = 1, 2. Note that the very last equation in the system of (7) is $\sum_{i\in s_1} p_i - \sum_{j\in s_2} q_j = 0$, this together with (6) imply that $\sum_{i\in s_1} p_i = 1$ and $\sum_{j\in s_2} q_j = 1$.

We can further rewrite (7) as

$$\sum_{i\in s_1} p_i\,\mathbf{u}_{1i} + \sum_{j\in s_2} q_j\,\mathbf{u}_{2j} = \mathbf{0}, \tag{8}$$

where

$$\mathbf{u}_{1i} = \begin{bmatrix} \mathbf{x}_{1i} \\ \mathbf{0} \\ \mathbf{z}_{1i} \end{bmatrix} - \frac{1}{2}\begin{bmatrix} \overline{\mathbf{X}}_1 \\ \overline{\mathbf{X}}_2 \\ \mathbf{0} \end{bmatrix}, \mathbf{u}_{2j} = \begin{bmatrix} \mathbf{0} \\ \mathbf{x}_{2j} \\ -\mathbf{z}_{2j} \end{bmatrix} - \frac{1}{2}\begin{bmatrix} \overline{\mathbf{X}}_1 \\ \overline{\mathbf{X}}_2 \\ \mathbf{0} \end{bmatrix}. \tag{9}$$

It is now clear that maximizing $l(\mathbf{p},\mathbf{q})$ under the restrictions (1), (2) and (3) is equivalent to maximizing $l(\mathbf{p},\mathbf{q})$ subject to (6) and (8). By using the Lagrange multiplier method we can show that

$$\hat{p}_i = \frac{d_{1i}^*}{1+\boldsymbol{\lambda}'\mathbf{u}_{1i}}, \hat{q}_j = \frac{d_{2j}^*}{1+\boldsymbol{\lambda}'\mathbf{u}_{2j}},$$

where $d_{ti}^* = 2d_{ti} / \left(\sum_{i\in s_1} d_{1i} + \sum_{j\in s_2} d_{2j}\right)$ for t = 1, 2, and the common Lagrange multiplier $\boldsymbol{\lambda}$ is the solution to

$$\sum_{t=1,2} \sum_{i\in s_t} \frac{d_{ti}^* \mathbf{u}_{ti}}{1+\boldsymbol{\lambda}'\mathbf{u}_{ti}} = \mathbf{0}. \tag{10}$$

The modified Newton-Raphson algorithm of Chen *et al.* (2002) can ideally be used to solve (10). Although such a modification is necessary for theoretical proof of convergence, it is our experience that the following conventional Newton-Raphson iteration procedure works well for almost all cases:

$$\boldsymbol{\lambda}^{(m+1)} = \boldsymbol{\lambda}^{(m)} + \left\{ \sum_{t=1,2} \sum_{i\in s_t} \frac{d_{ti}^* \mathbf{u}_{ti} \mathbf{u}_{ti}'}{\left(1+\left[\boldsymbol{\lambda}^{(m)}\right]'\mathbf{u}_{ti}\right)^2} \right\}^{-1} \sum_{t=1,2} \sum_{i\in s_t} \frac{d_{ti}^* \mathbf{u}_{ti}}{1+\left[\boldsymbol{\lambda}^{(m)}\right]'\mathbf{u}_{ti}},$$

with the initial value of $\boldsymbol{\lambda}$ chosen as $\mathbf{0}$.

This second algorithm is applicable under general situations. It requires to solve (10) only once using the existing well-behaved algorithm of Chen *et al.* (2002) and can be programmed by survey users using popular statistical software such as SAS or R/Splus.

2.3 A comparison to Zieschang's regression method

The combined empirical likelihood approach proposed in this article has the same spirit of the composite generalized regression estimator proposed by Zieschang (1990). This is evident when we compare the constraints (7) used here to the enlarged regression system (3.10) used by Zieschang. There are several advantages, however, from using the empirical likelihood method. In addition to its clear maximum likelihood interpretations, the EL estimator is computed based on the normalized intrinsic positive weights, i.e., $\hat{p}_i > 0$ and $\sum_{i\in s_1} \hat{p}_i = 1$. This latter feature is particularly appealing to the potential users of the survey data since the published weights are often used for a variety of purposes, including the estimation of proportions or more generally the finite population distribution function $F(y)$. The EL estimator $\hat{F}_{EL}(y)$ will be itself a genuine distribution function. It is range-respecting and can be inverted to get quantile estimates.

An explicit relationship between the EL estimator and a generalized regression-type estimator can be established.

Theorem 1. *Under suitable regularity conditions, the maximum pseudo empirical likelihood estimators* $\hat{\bar{Y}}_1 = \sum_{i\in s_1} \hat{p}_i y_{1i}$ *and* $\hat{\bar{Y}}_2 = \sum_{j\in s_2} \hat{q}_j y_{2j}$ *are asymptotically equivalent to a generalized regression-type estimator, i.e.,*

$$\hat{\bar{Y}}_t = \bar{y}_t + \hat{\mathbf{B}}_{t1}'\left(\bar{\mathbf{X}}_1 - \bar{\mathbf{x}}_1\right) + \hat{\mathbf{B}}_{t2}'\left(\bar{\mathbf{X}}_2 - \bar{\mathbf{x}}_2\right) + \hat{\mathbf{B}}_{t3}'\left(\bar{\mathbf{z}}_2 - \bar{\mathbf{z}}_1\right) + o_p\left(n^{-1/2}\right), \tag{11}$$

where $\bar{y}_t = \sum_{i\in s_t} d_{ti}^* y_{ti}, \bar{\mathbf{x}}_t = \sum_{i\in s_t} d_{ti}^* \mathbf{x}_{ti}, \bar{\mathbf{z}}_t = \sum_{i\in s_t} d_{ti}^* \mathbf{z}_{ti}, n = n_1 + n_2$, *and the combined "regression coefficients"* $\hat{\mathbf{B}}_t = \left(\hat{\mathbf{B}}_{ti}', \hat{\mathbf{B}}_{t2}', \hat{\mathbf{B}}_{t3}'\right)'$ *are computed as*

$$\hat{\mathbf{B}}_t = \left(\sum_{t=1,2} \sum_{i \in s_t} d_{ti}^* \, \mathbf{u}_{ti} \, \mathbf{u}_{ti}' \right)^{-1} \sum_{i \in s_t} d_{ti}^* \, \mathbf{u}_{ti} \, y_{ti} \, ,$$

with the $\mathbf{u}_{ti}$ *defined by (9).*

The required regularity conditions and proofs of the theorems can be found in a technical report available from the author. Note that the combined auxiliary information $\overline{\mathbf{X}}_1, \overline{\mathbf{X}}_2, \overline{\mathbf{z}}_1$ and $\overline{\mathbf{z}}_2$, as well as the basic design weights d_{1i} and d_{2j} from both surveys all appeared explicitly in the equivalent generalized regression estimator, an estimator that is quite unique from the conventional point of view. Further, if both sampling designs satisfy $\sum_{i \in s_1} d_{1i} = \sum_{i \in s_2} d_{2i} = N$, as is the case under simple random sampling or stratified random sampling, then $d_{ti}^* = d_{ti} / N$, the estimators $\bar{y}_t, \overline{\mathbf{x}}_t$ and $\overline{\mathbf{z}}_t$ all reduce to the usual Horvitz-Thompson estimators for the corresponding population means.

3. THE SEPARATE EL APPROACH

In the combined approach the unknown population mean vector $\overline{\mathbf{Z}}$ is implicitly estimated by the maximum pseudo empirical likelihood estimator $\hat{\boldsymbol{\theta}}$ from the pooled sample. This can be seen from the first algorithm presented in Section 2.1. Some computational complications arising from the combined approach are solely due to the attempt in estimating $\overline{\mathbf{Z}}$ by $\hat{\boldsymbol{\theta}}$.

One way to circumvent this difficulty is to take a two-step approach. Suppose we replace $\hat{\boldsymbol{\theta}}$ by a different estimator of $\overline{\mathbf{Z}}$, say $\overline{\mathbf{z}}$, using the combined data from both surveys. We then use this $\overline{\mathbf{z}}$ as control values for the constraints used in the empirical likelihood estimation for each of the two surveys. By doing so we not only bring consistency for the common auxiliary variables ***z*** between the two surveys but also improve the resulting estimators $\hat{\bar{Y}}_1$ and $\hat{\bar{Y}}_2$ if $\overline{\mathbf{z}}$ is suitably constructed from the combined sample data. This is similar to the case of two-phase sampling where the unknown population quantity $\overline{\mathbf{Z}}$ is estimated using the large first phase sample.

Estimation of $\bar{Y}_1$ and $\bar{Y}_2$ using a pre-determined $\overline{\mathbf{z}}$ as control value becomes two separate estimation problems. For instance, the EL estimator for $\bar{Y}_1$ is given by $\hat{\bar{Y}}_1 = \sum_{i \in s_1} \hat{p}_1 y_{1i}$, where the $\hat{p}_i$ maximize $l(\mathbf{p}) = \sum_{i \in s_1} d_{1i} \log(p_i)$ subject to constraints

$$\sum_{i \in s_1} p_i = 1 (p_i > 0), \sum_{i \in s_1} p_i \mathbf{x}_{1i} = \overline{\mathbf{X}}_1 \text{ and } \sum_{i \in s_1} p_i \mathbf{z}_{1i} = \overline{\mathbf{z}} \, .$$

The resulting weights are computed as $\hat{p}_i = d_{1i}^* / (1 + \boldsymbol{\lambda}' \mathbf{u}_{1i})$, where $d_{1i}^* = d_{1i} / \sum_{i \in s_1} d_{1i}$ and the Lagrange multiplier $\boldsymbol{\lambda}$ is the solution to

$$\sum_{i \in s_1} \frac{d_{1i}^* \, \mathbf{u}_{1i}}{1 + \boldsymbol{\lambda}' \mathbf{u}_{1i}} = 0 \text{, with } \mathbf{u}_{1i} = \begin{pmatrix} \mathbf{x}_{1i} - \overline{\mathbf{X}}_1 \\ \mathbf{z}_{1i} - \overline{\mathbf{z}} \end{pmatrix}. \tag{12}$$

The algorithm of Chen *et al.* (2002) can directly be used here to obtain $\boldsymbol{\lambda}$ without any modification.

The major issue under this separate EL approach is the choice of $\overline{\mathbf{z}}$. Renssen and Nieuwenbroek (1997) provided an excellent account on the estimation of $\overline{\mathbf{Z}}$ using combined sample data. They suggested a general class of estimators in the form of $\overline{\mathbf{z}} = \mathbf{P} \, \overline{\mathbf{z}}_1 + \mathbf{Q} \, \overline{\mathbf{z}}_2$, where ***P*** and ***Q*** are two matrices with compatible dimension to ***z*** such that

$\boldsymbol{P} + \boldsymbol{Q} = \boldsymbol{I}$, and $\bar{\mathbf{z}}_t$ is the generalized regression estimator of $\boldsymbol{Z}$ with $\mathbf{x}_t$ as auxiliary variables. In the absence of $\mathbf{X}_t$, one can take $\bar{\mathbf{z}}_t$ as the Horvitz-Thompson estimator of $\bar{\mathbf{Z}}$ using data from the tth survey.

Two choices of the matrix pair ($\boldsymbol{P}$, $\boldsymbol{Q}$) will be examined in the simulation study presented in the next section. The simplest one is the proportional combination where $\mathbf{P} = (n_1 + n_2)^{-1} n_1 \mathbf{I}$ and $\mathbf{Q} = (n_1 + n_2)^{-1} n_2 \mathbf{I}$; the optimal combination uses

$$\mathbf{P} = V_p(\bar{\mathbf{z}}_2)\left[V_p(\bar{\mathbf{z}}_1) + V_p(\bar{\mathbf{z}}_2)\right]^{-1} \text{ and } \mathbf{Q} = V_p(\bar{\mathbf{z}}_1)\left[V_p(\bar{\mathbf{z}}_1) + V_p(\bar{\mathbf{z}}_2)\right]^{-1},$$

where $V_p(\bar{\mathbf{z}}_t)$ is the design-based variance-covariance matrix of $\bar{\mathbf{z}}_t$. Note that the $\boldsymbol{P}$ used in the optimal combination can also be written as $\mathbf{P} = \left\{\left[V_p(\bar{\mathbf{z}}_1)\right]^{-1} + \left[V_p(\bar{\mathbf{z}}_2)\right]^{-1}\right\}^{-1}\left[V_p(\bar{\mathbf{z}}_1)\right]^{-1}$, and similarly for $\boldsymbol{Q}$ as well. This choice is optimal since it minimizes $V_p(\mathbf{a}'\bar{\mathbf{z}})$ for an arbitrary constant vector $\boldsymbol{a}$ among the general class of estimators considered by Renssen and Nieuwenbroek (1997). When simple random sampling is used for both surveys and $\bar{\mathbf{z}}_t$ are the simple sample means, the optimal combination reduces to the proportional one if the two sampling fractions are the same or can be ignored. Note that for the optimal combination the matrices $\boldsymbol{P}$ and $\boldsymbol{Q}$ need to be replaced by sample-based estimates for applications.

The separate EL approach is less elegant than the combined one in terms of maximum likelihood estimation. This approach, however, is intuitively attractive, and computation in this case is straightforward and simple. Under suitable regularity conditions similar to those used in Theorem 1, we can show that the separate EL estimator is asymptotically equivalent to the regression estimator discussed by Renssen and Nieuwenbroek (1997), i.e.,

$$\hat{\bar{\mathbf{Y}}}_t = \bar{y}_t + \hat{\mathbf{B}}'_{t1}(\mathbf{X}_t - \bar{\mathbf{x}}_t) + \hat{\mathbf{B}}'_{t2}\left(\bar{\mathbf{z}} - \bar{\mathbf{z}}_t^*\right) + o_p\left(n^{-1/2}\right), \tag{13}$$

where $\bar{y}_t = \sum_{i \in s_t} d_{ti}^* y_{ti}$, $\bar{\mathbf{x}}_t = \sum_{i \in s_t} d_{ti}^* \mathbf{x}_{ti}$, $\bar{\mathbf{z}}_t^* = \sum_{i \in s_t} d_{ti}^* \mathbf{z}_{ti}$, $d_{ti}^* = d_{ti} / \sum_{i \in s_t} d_{ti}$, and the regression coefficients $\hat{\mathbf{B}}_t = \left(\hat{\mathbf{B}}'_{t1}, \hat{\mathbf{B}}'_{t2}\right)'$ are given by

$$\hat{\mathbf{B}}_t = \left(\sum_{i \in s_t} d_{ti}^* \mathbf{u}_{ti} \mathbf{u}'_{ti}\right)^{-1} \sum_{i \in s_t} d_{ti}^* \mathbf{u}_{ti} y_{ti},$$

where $\mathbf{u}_{1i}$ (and $\mathbf{u}_{2i}$ in obvious form) are defined in (12). The first two terms on the right hand side of (13) can be viewed as a generalized regression estimator for $\bar{Y}_t$ based on auxiliary variables $\mathbf{x}_t$, and the third one is an adjusting term trying to further improve the regression estimator using the extra information on $\boldsymbol{z}$ variables.

It is worthwhile to note that for the separate EL estimator the different sample sizes can easily be taken into account for the estimation of $\bar{\mathbf{Z}}$. The combined EL approach, however, does not automatically accommodate this and requires a special weighting adjustment to achieve the same goal. Further development over this direction will not be pursued here but this argument provides a possible explanation why the combined EL estimator is often outperformed by the separate one as seen from the simulation results reported in the next section.

4. SIMULATION STUDY

In this section we examine the finite sample performance of the proposed estimators through a limited simulation study. The finite population used in this study was based on the real data from the 1996 Statistics Canada's Family Expenditure (FAMEX) Survey for the province of Ontario. The data set contains N=2396 observations measured

over a variety of characteristics. Variables which are relevant to our study include x_1: number of children (age <15); x_2: number of youths (age 15-24); x_3: number of people in the household; *z*: total income after taxes; and *y*: total expenditure.

In the simulation, we treat the data set itself as a finite population. This population is further split into eight strata according to the original sampling design. For the first survey, the number of children (x_1) and the number of people (x_3) with known population means are used as control variables and the total expenditure *y* is treated as the response variable. We also assume that the variables x_2 and x_3 are used as control variables in the second survey, and information on total income (*z*) is conveniently collected for both surveys but the population mean $\bar{Z}$ is unknown. The goal is to estimate the population mean $\bar{Y}$ using all useful information while respecting the consistency requirement imposed over the *z* variable for the two surveys.

For each simulation run, a stratified random sample of size n_t under proportional allocation is taken for the *t*th survey, *t* = 1, 2, and three maximum pseudo empirical likelihood estimators for $\bar{Y}$ are computed. Let EL(C) denote the combined EL estimator, EL(SP) be the separate EL estimator using proportional combination for the estimation of $\bar{Z}$, and EL(SO) represent the separate EL estimator using optimal combination in estimating $\bar{Z}$. Also computed for each simulation are three GREG-type estimators: the one proposed by Zieschang (1990) is denote by GR(Z), which is equivalent to our combined EL estimator EL(C); and the estimators proposed by Renssen and Nieuwenbroek (1997) are denoted by GR(RN1) and GR(RN2), corresponding to our EL(SP) and EL(SO), respectively. The Λ matrix used to formulate the GR(Z) estimator is taken as $diag(d_1,...d_n)$. The process is repeated independently for *B* = 1000 times.

The performance of an estimator $\hat{\bar{Y}}$ is measured in terms of the simulated Relative Bias (RB) and Relative Efficiency (RE) defined as

$$RB = \frac{1}{B}\sum_{b=1}^{B}\frac{\hat{\bar{Y}}(b)-\bar{Y}}{\bar{Y}} \quad \text{and} \quad RE = \frac{MSE\left(\hat{\bar{Y}}_0\right)}{MSE\left(\hat{\bar{Y}}\right)},$$

where $\hat{\bar{Y}}(b)$ is the estimator $\hat{\bar{Y}}$ computed from the *b*th simulated sample, $MSE(\hat{\bar{Y}}) = B^{-1} \sum_{b=1}^{B}\left[\hat{\bar{Y}}(b)-\bar{Y}\right]^2$, and $\hat{\bar{Y}}_0$ is the baseline estimator for comparison. In our study $\hat{\bar{Y}}_0$ is chosen as the generalized regression estimator (GREG) of $\bar{Y}$ using x_1 and x_3 as auxiliary variables. Note that the sample information on *z* cannot be used here for the regression estimation of $\bar{Y}$ since $\bar{Z}$ is assumed unknown. The sample sizes n_t = 80, 160 and 240 used in the simulation represent a typical sampling fraction at 2.5%, 5% and 10% level.

Table 1: Simulated relative efficiencies based on the 1996 Statcan FAMEX survey data

n_1	n_2	GREG	EL(C)	GR(Z)	EL(SP)	GR(RN1)	EL(SO)	GR(RN2)
80	80	1.00	1.19	1.13	1.28	1.28	1.28	1.28
	160	1.00	1.36	1.28	1.31	1.43	1.31	1.42
	240	1.00	1.42	1.31	1.43	1.49	1.48	1.49
160	80	1.00	1.03	0.91	1.15	1.15	1.15	1.15
	160	1.00	1.28	1.10	1.27	1.26	1.27	1.27
	240	1.00	1.33	1.18	1.29	1.28	1.30	1.30
240	80	1.00	0.91	0.80	1.16	1.14	1.15	1.13
	160	1.00	1.11	0.95	1.16	1.14	1.16	1.15
	240	1.00	1.24	1.07	1.22	1.20	1.24	1.23

The absolute values of the simulated relative biases are all less than 0.1% and are not reported here. Table 1 reports the relative efficiency of the EL and the GREG-type estimators under various scenarios of the sample size combinations. Our major findings can be summarized as follows:

(i) the two separate EL estimators have similar performance and they both perform well. The gain of efficiency is more pronounced when the second sample size is larger;

(ii) the combined EL estimator has satisfactory performance when the second sample has a compatible size but could have deteriorated performance otherwise (i.e., the case of n_1 =240 and n_2 =80);

(iii) the GREG estimators of Renssen and Nieuwenbroek (1997) have very similar performance to the separate EL estimators, but the GREG estimator of Zieschang (1990) is outperformed by the combined EL estimator at all cases;

(iv) the use of information on the common auxiliary variable *z* provides substantial improvement over the baseline GREG estimator when the second sample is not too small.

5. CONCLUDING REMARKS

Adjusting weights to satisfy certain efficiency and consistency requirements is a constant theme in survey sampling, and having positively adjusted weights is a highly desirable property for the users of production micro data files where the weights are viewed as the number of units in the finite population represented by the sampled unit. Positive weights will also guarantee positive estimation for known positive population quantities.

It should be noted that positive weights in regression estimation can theoretically be achieved through constrained minimization under the context of calibration estimation as discussed in Deville and Särndal (1992). Practical implementation of such a method, however, is not straightforward and often involves ad hoc approximations. The loss of efficiency due to these approximations is usually unknown. The empirical likelihood method, on the other hand, provides a natural way of doing this with the final adjusted weights that are intrinsically positive. It should also be noted that the total amount of time required for computing the proposed EL estimators remains limited. In our simulation study, it takes less than 20 seconds on a dual process Sun Unix workstation to compute the combined EL estimator when $n_1 = n_2 = 240$ and the program is written in R/Splus.

While the adjusted weights using a GREG-type technique tend to have some small or negative values, the weights obtained from the EL approach can occasionally contain a few large values. For the simulation results reported in Section 4 where a proportional sample size allocation scheme is used to draw the two stratified random samples, the g-weights given by $g_i = w_i / d_i$ are all within the range of (0.25, 4.00), where w_i denote the EL adjusted weights and d_i represent the basic design weights. More than 99% of these g-weights are indeed between 0.50 and 2.00. If we use an unbalanced allocation scheme where the largest stratum $(N_h = 763)$ and the smallest one $(N_h = 33)$ receive equal sample sizes, we observed that a couple of g-weights can be larger than 4.00 or even 6.00. Theoretically this is not a problem regarding the statistical properties of the EL estimators. For users who also have concerns about large weights, the idea of minimum relaxation of constraints presented in Chen *et al.* (2002) can be applied to obtain more general range restricted weights through the EL method.

It is easy to argue, both theoretically and empirically, that under ideal situations the gain of efficiency from using the combined sample data is almost guaranteed. Cases where the forced consistency requirement over the common variables will likely detriment the resulting estimators may include (1) severe uncontrolled nonsampling errors; (2) extremely unbalanced sampling designs or sample size allocations; (3) misconceptualized target populations; (4) weak or even no correlation between the common variables and the response variables; and (5) the use of questionable common variables.

The successful use of the proposed EL method for combining information on common auxiliary variables for real surveys requires detailed considerations at the planning stage and careful discretions at the estimation stage. As pointed out by Renssen and Nieuwenbroek (1997), common variables in the strict sense are not easily found due to discrepancies between definitions, methods of observation, and reference periods. Such complications can be reduced if the involved surveys are harmonized at the design stage. For example, in split questionnaire design

where certain common questions are contained in both versions of the questionnaire, attention should be given to the ordering and positioning of these common questions to reduce the potential response bias and/or carry-over effect. For human population surveys conducted regularly over time, variables related to gender, age, educational background, etc, can easily be conceived as common variables if the change of population dynamics over certain time periods can be ignored. Other variables such as "Employment Status in May 2003", which can be measured "directly" during a June 2003 survey but require "recalled" answers for surveys conducted in a later time, need to be treated with care. Only if the "recalled" answers are as accurate as the "direct" measurement can this type of variables be treated as common variables.

ACKNOWLEDGEMENTS

This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada. Comments and suggestions from Professor Jiahua Chen are gratefully acknowledged.

REFERENCES

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, 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. *Journal of the American Statistical Association*, 87, 376-382.

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

Isaki, C.T. and Fuller, W.A. (1982). Survey Design Under the Regression Superpopulation Model. *Journal of the American Statistical Association*, 77, 89-96.

Owen, A.B. (1990). Empirical Likelihood Ratio Confidence Regions. *Annals of Statistics*, 18, 90-120.

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

Renssen, R.H. and Nieuwenbroek, N.J. (1997). Aligning Estimates for Common Variables in Two or More Sample Surveys. *Journal of the American Statistical Association*, 92, 368-374.

Zieschang, K.D. (1990). Sample Weighting Methods and Estimation of Totals in the Consumer Expenditure Survey. *Journal of the American Statistical Association*, 85, 986-1001.