AISM (2006) 58: 499–509
DOI 10.1007/s10463-005-0013-9

Changchun Wu · Runchu Zhang

An information-theoretic approach to the effective usage of auxiliary information from survey data

Received: 24 May 2004 / Revised: 4 March 2005 / Published online: 28 June 2006

Abstract In this paper, we propose an information-theoretic approach to the effective usage of auxiliary information from survey data, which is suitable for both simple and complex survey data. Our estimator under simple random sampling without replacement will be consistent and asymptotically normal. We show that the resulting estimates have smaller asymptotic variances than the usual estimates which do not use auxiliary information. For more complex survey designs, the resulting estimator is in essence asymptotically equivalent to a pseudo empirical likelihood estimator. Results of a limited simulation study show that the proposed estimators perform well among a number of competitors.

Keywords Calibration · Entropy · Cross-entropy · Generalized regression estimator · Empirical likelihood · Optimal regression estimator · Jackknife

1 Introduction

In sample surveys, auxiliary information on the finite population is regularly used to increase the precision of estimators, most commonly estimators of the population mean or total. Ratio and regression estimators incorporate known finite population means of auxiliary variables. Calibration estimators adjust basic survey weights so the sample sum of a weighted auxiliary variable equals its known population totals (see Deville and Särndal, 1992). However, using the methods above,

C. Wu (✉)
School of Mathematics and Information
Jiaxing University, Jiaxing, Zhejiang, 314001, China
E-mail: hnwucc@eyou.com

C. Wu · R. Zhang
LPMC and School of Mathematical Sciences
Nankai University,
Tianjin, 300071, China

the resulting weights may not be always positive. Chen and Qin (1993) proposed an empirical likelihood approach to the usage of auxiliary information in simple random sampling without replacement. Their theoretical and simulation results suggest the approach has desirable properties when estimating means, totals and population distribution functions, as well as quantiles. Unfortunately, their formulation of the method does not extend to more complex survey designs. Chen and Qin (1999) developed a pseudo-empirical likelihood approach for complex surveys which reduces to Chen and Qin's method in the case of simple random sampling. They showed that the method is asymptotically equivalent to a generalized regression estimator in the case of estimating a mean or population distribution function with known population means for a vector of auxiliary variables. However, from the definitions (Chen and Qin, 1993), we know that the defined empirical likelihood function is not the true likelihood function, and that it is only an approximation or a design unbiased estimate of the true likelihood function when the entire finite population is viewed as independent and identically distributed sample from some super-population. Although empirical likelihood approach has desirable large sample properties, it may not be the best method.

In this paper, we develop a cross-entropy minimization (CEM) approach to the effective usage of auxiliary information, which is suitable for both simple and complex survey data. We argue that the cross-entropy minimization estimator (CEM) be more appealing than the empirical maximum likelihood (EML) or pseudo empirical maximum likelihood (PEML) estimator which has been the focus of most research. The first reason concerns the interpretation of both estimators as minimizing a (directed) distance between the estimated probabilities π_i and the empirical frequencies $1/n$ or d_i/N. It seems appealing to weight the discrepancies using an efficient estimate of these probabilities (i.e.,$\hat{\pi}_i$), as in CEM procedure, rather than by an inefficient estimate of these probabilities (i.e., $1/n$ or d_i/N), as in the EML or PEML procedure. The second reason concerns the relative robustness of the two estimators. The CEM estimator is affected to a much lesser extent by perturbations in data (see Imbens, Spady, and Johnson, Imbens et al. 1998, Hellerstein and Imbens, 1999). Our method is a computationally simple approach. In Sect. 2, we give a brief review of entropy and a justification for its use in finite populations under simple random sampling, we show that the limiting distribution of CEME $\hat{\theta}_n$ is identical with the EMLE in Chen and Qin (1993). In Sect. 3, we extend the method to more complex survey designs, the resulting estimator is in essence asymptotically equivalent to a pseudo empirical likelihood estimator. In particular, our estimator is asymptotically equivalent to a generalized regression estimator in case of estimating a mean or population distribution function with known population means for a vector of auxiliary variables. In Sect. 4, a limited simulation is conducted to study the finite sample properties of the proposed estimators. Simulation results show that the CEMEs perform well among a number of competitors. Some proofs are given in Appendix.

2 Entropy, cross-entropy, estimation under simple random sampling

The importance of suitable measures of distance between probability distributions arises because of the role they play in the problems of inference and discrimination. The concept of distance between two probability distributions was initially developed by Mahalanobis (See Bero and Bilias, 2002). Since then various types of distance measures have been developed in the literature. Here we consider Shannon's concept of information-theoretic entropy and its generalization known as the Kullback and Leibler relative entropy (i.e., the cross entropy) or the divergence measure between two probability distributions. Any probability distribution $p_i, i = 1, 2, \ldots, n$ (say) of a random variable taking n values provides a measure of uncertainty regarding that random. In the information theory literature, this measure of uncertainty is called entropy. Entropy is generally taken as a measure of expected information, that is how much information do we have in the probability distribution $p_i, i = 1, 2, \ldots, n$. Intuitively, information should be a decreasing function of p_i, i.e., the more unlikely an event, the more interesting it is to know that it can happen. A simple choice for such a function is $-\log p_i$. Entropy $H(\mathbf{p})$ is defined as a weighted sum of the information $-\log p_i, i = 1, 2, \ldots, n$ with respective probabilities as weights, namely,

$$H(\mathbf{p}) = -\sum_{i=1}^{n} p_i \log p_i. \tag{1}$$

If $p_i = 0$ for some i, the $p_i \log p_i$ is taken to be zero.

Following Eq. (1), the cross-entropy of one probability distribution $\mathbf{p} = (p_1, p_2, \ldots, p_n)'$ with respect to another distribution $\mathbf{q} = (q_1, q_2, \ldots, q_n)'$ can be defined as

$$C(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{n} p_i \log(\frac{p_i}{q_i}), \tag{2}$$

which is a measure of distance between two distributions. If we choose $\mathbf{q} = (1/n, 1/n, \ldots, 1/n)' = \mathbf{1}/n$,where $\mathbf{1}$ is a $n \times 1$ vector of ones, $C(\mathbf{p}, \mathbf{q})$ reduces to

$$C\left(\mathbf{p}, \frac{\mathbf{1}}{n}\right) = \sum_{i=1}^{n} p_i \log p_i - \log n. \tag{3}$$

Therefore, entropy maximization is a special case of CEM with respect to the uniform distribution. If we try to find a probability distribution that maximizes the entropy $H(\mathbf{p})$ in Eq. (1) (or minimizes the cross-entropy $C(\mathbf{p}, \mathbf{1}/n)$in Eq. 3), the optimal solution is the uniform distribution, i.e., $\mathbf{p} = \mathbf{1}/n$. From what said above, the cross-entropy can then be used as a objective function in the estimation of finite population parameters.

Suppose a finite population, $\mathcal{S}$, consists of N distinct units with measurements $z_i = (y_i, x_i), i = 1, \ldots, N$; s, of size n, is a sample from $\mathcal{S}$. For the purpose of illustration, let us first consider simple random sampling without replacement. Most finite population parameters in surveys can be expressed as a functional $\theta_N = \theta(F_N)$, such as the finite population mean $\bar{Y} = N^{-1}\sum_{i=1}^{N} y_i$, or the finite population distribution function $F_N = N^{-1}\sum_{i=1}^{N} \delta_{z_i}$ itself, where δ_{z_i} is the point measure at z_i. Furthermore, auxiliary information can usually be expressed in the form $E\{u(z)\} = N^{-1}\sum_{i=1}^{N} u(z_i) = 0$ for some function u. For example, $u(z_i) = x_i - \bar{X}_N, i=1, \ldots, N$, where $\bar{X}_N = N^{-1}\sum_{i=1}^{N} x_i$ is the auxiliary variable finite population mean. If there is no auxiliary information, the optimal estimator of F_N is the well-known empirical distribution function $F_n = 1/n \sum_{i\in s} \delta_{z_i}$. When we have auxiliary information $E\{u(z)\}=0$, our objective is to derive the new estimator of F_N that modify as little as possible F_n, which then reduces to minimize the cross-entropy $C(\mathbf{p}, \mathbf{1}/n)$ under constraint $E\{u(z)\} = 0$. As argued by Owen (1990), we need consider only estimates of F_N whose support is contained in the set of observations. The problem becomes to: minimize

$$C\left(\mathbf{p}, \frac{\mathbf{1}}{n}\right)$$

subject to

$$\sum_{i\in s} p_i = 1, \quad \sum_{i\in s} p_i u(z_i) = 0, \qquad (0 \le p_i \le 1). \tag{4}$$

Using the Lagrange multiplier method, we find the solution of Eq. (4) satisfies

$$\hat{p}_i = \frac{\exp\{\lambda u(z_i)\}}{\sum_{j\in s} \exp\{\lambda u(z_j)\}}, \quad i \in s, \tag{5}$$

where λ satisfies the following equation

$$\sum_{i\in s} u(z_i) \exp\{\lambda u(z_i)\} = 0. \tag{6}$$

For any parameters that can be written as $\theta_N = \theta(F_N)$, the resulting CEME is given by

$$\hat{\theta}_n = \theta(\hat{F}_n), \quad \hat{F}_n = \sum_{i\in s} \hat{p}_i \delta_{z_i}.$$

In particular, if $\theta_N = E_N\{g(y)\} = N^{-1} \sum_{i=1}^N g(y_i)$, then

$$\hat{\theta}_n = \frac{\sum_{i\in s} g(y_i) \exp\{\lambda u(z_i)\}}{\sum_{i\in s} \exp\{\lambda u(z_i)\}},$$

with λ satisfying Eq. (6).

To study the asymptotic properties of the CEMEs, we assume simple random sampling without replacement where both the sample size n and the finite population size N go to infinity as a certain index ν goes to infinity. However, for convenience, we will suppress the index ν in the following whenever possible, and let $g_i = g(y_i)$, $u_i = u(z_i)$, and define

$$\bar{g} = N^{-1} \sum_{i=1}^N g_i, \quad \sigma_g^2 = (N-1)^{-1} \sum_{i=1}^N \{g_i - \bar{g}\}^2,$$

$$\sigma_u^2 = (N-1)^{-1} \sum_{i=1}^N u_i^2, \quad \sigma_{gu} = (N-1)^{-1} \sum_{i=1}^N \{g_i - \bar{g}\} u_i.$$

The CEM estimator $\hat{\theta}_n$ under simple random sampling without replacement will be consistent and asymptotically normal.

Theorem 2.1 *Suppose that as* $\nu \longrightarrow \infty$, *the population size* N, *sample size* n, *and* $N-n$ *go to infinity, and*

$$N^{-1}\sum_{i=1}^{N}|u_i|^3, \quad N^{-1}\sum_{i=1}^{N}|g_i|^3$$

have an upper bound independent of ν. *Then*

$$\frac{\sqrt{n}(\hat{\theta}_n-\theta)}{\sigma_\nu} \xrightarrow{d} N(0,1),$$

where "$\xrightarrow{d}$" *denotes convergence in distribution, and* $\sigma_\nu^2 = (1-(n/N))(\sigma_g^2 - \sigma_{gu}^2/\sigma_u^2)$.

Remark 2.1 This result shows that, whenever there is any auxiliary information, the asymptotic unconditional variance of $n^{\frac{1}{2}}(1-n/N)^{-\frac{1}{2}}(\hat{\theta}_n-\theta)$ with respect to simple random sampling is always smaller than or equal to σ_g^2. We also note that the reduction of the asymptotic variance depends on the relevance of the auxiliary information. The larger the correlation between $u(z)$ and $g(y)$, the greater the gain in precision.

Remark 2.2 The asymptotic efficiency of our method is equivalent to that of the regression method, or to that of using the optimal estimating equation. Unlike the regression method, our method does not specify a model. Also, our method does not estimate the parameter

$$A_N = \text{cov}_N\frac{\{g(Y),u(Z)\}}{\text{var}_N\{g(Y)\}},$$

which is needed in the optimal estimating equation (Godambe and Thompson, 1986).

There can be many ways to estimating σ_ν^2 such as estimating σ_g^2, σ_{gu}, and σ_u^2 separately. However, one might apply re-sampling variance estimators such as the jackknife, bootstrap (see Chen and Qin, 1993) directly to $\hat{\theta}_n$, recalculating the $\hat{p}_i$ for each re-sample. These may perform better for finite samples since they are applied directly to $\hat{\theta}_n$. Here we only consider the jackknife . A result on its large sample property is given in the following theorem.

Theorem 2.2 *Under the same conditions of Theorem 2.1, let* $\hat{\theta}_{-j}$ *be the estimator when the* j*th observation is removed and define*

$$\hat{\sigma}_\mathrm{J}^2 = \left(1-\frac{n}{N}\right)(n-1)\sum_{j\in s}(\hat{\theta}_n-\hat{\theta}_{-j})^2,$$

be the jackknife variance estimator. Then $\hat{\sigma}_\mathrm{J}^2$ *is consistent.*

3 Estimation under complex sampling

In this section , we extend the CEM method to more complex sampling . Consider a finite population , $\mathcal{S}$, of N distinct units with measurements $z_i = (y_i, x_i)$ as in the previous section. But now suppose the sample, s, is drawn using some sampling design, $p(\cdot)$, that is, the sample $s \subset \mathcal{S}$ is drawn with probability $p(s)$. Assume the inclusion probabilities $\pi_i = p_r(i \in s)$ are strictly positive. When no auxiliary information can be used, the conventional estimator of F_N is the Horvitz-Thompson estimator $\hat{F}_{\text{HT}} = N^{-1} \sum_{i \in s} d_i \delta_{z_i}$, where $d_i = 1/\pi_i$ are called the basic design weights. Now we have auxiliary information such as $E_N\{u(z)\} = 0$. Of course, the estimator should be sought among distribution functions satisfying $E_N\{u(z)\} = 0$ and modified as little as possible the Horvitz-Thompson estimator $\hat{F}_{\text{HT}} = N^{-1} \sum_{i \in s} d_i \delta_{z_i}$. The problem then becomes to: minimize

$$C\left(\mathbf{p}, \frac{\mathbf{d}}{N}\right) = \sum_{i \in s} p_i \log\left(N \frac{p_i}{d_i}\right) \tag{7}$$

subject to

$$\sum_{i \in s} p_i = 1, \quad \sum_{i \in s} p_i u_i = 0, \qquad (0 \le p_i \le 1).$$

Using the Lagrange multiplier method, it is easily shown that, for any parameters that can be written as $\theta_N = \theta(F_N)$, the resulting CEME is given by $\hat{\theta}_n = \theta(\hat{F}_n)$, where $\hat{F}_n = \sum_{i \in s} \hat{p}_i \delta_{z_i}$, with

$$\hat{p}_i = \frac{d_i \exp\{\lambda u_i\}}{\sum_{j \in s} d_j \exp\{\lambda u_j\}}, \quad i \in s \tag{8}$$

and the Lagrange multiplier, λ, is the solution to

$$\sum_{i \in s} d_i u_i \exp\{\lambda u_i\} = 0. \tag{9}$$

If $u(\cdot)$ is vector valued, this extends naturally using a vector valued λ.

Now we study the asymptotic properties of the CEM estimators. Assume that there is a sequence of finite populations indexed by ν such that when $\nu \to \infty$:

(i) $N, n, N - n$ go to infinity;

(ii) $N^{-1} \sum_{i=1}^{N} |u_i|^3 = O_p(1), \qquad N^{-1} \sum_{i=1}^{N} |g_i|^3 = O_p(1)$;

(iii) $u^* = \max_{i \in s} |u_i| = o_p(n^{\frac{1}{2}})$;

(iv) $\sum_{i \in s} d_i u_i / \sum_{i \in s} d_i u_i^2 = O_p(n^{-\frac{1}{2}})$;

(v) $0 < c_1 \le \frac{S_{wu}^2}{\sigma_u^2} \le c_2 < \infty$, where $S_{wu}^2 = \sum_{i \in s} w_i u_i^2, \qquad w_i = \frac{d_i}{\sum_{j \in s} d_j}$.

We have stated these necessary conditions in a general form which is compact but not enlightening. Many commonly used sampling designs satisfy these conditions under some moderate assumptions, but do not involve stratification.

Theorem 3.1 *Under the conditions (i)–(v) above, we have*

$$\hat{\theta}_n = \bar{g}_w - \frac{S_{wug}}{S^2_{wu}} \bar{u}_w + o_p(n^{-\frac{1}{2}}),$$

where $\bar{g}_w = \sum_{i \in s} w_i g_i$, $\bar{u}_w = \sum_{i \in s} w_i u_i$, *and* $S_{wug} = \sum_{i \in s} w_i (u_i - \bar{u}_w)(g_i - \bar{g}_w)$.

Hartley and Rao (1968) consider the problem of estimating the population mean $\bar{Y}_N$ when $\bar{X}_N$ is a known scalar in the case of simple random sampling without replacement, and showed that maximizing the empirical likelihood is asymptotically equivalent to a regression estimator. In this more general setting, Chen and Qin (1999) obtain a similar result. From Theorem 3.1, we also have

Corollary 3.1 *Under conditions (iii) and (iv) above, the CEME* $\hat{\bar{Y}}_N$ *of* $\bar{Y}_N$, *when* $\bar{X}_N$ *is known, is asymptotically equivalent to a generalized regression estimator (GREG). That is,*

$$\hat{\bar{Y}}_N = \sum_{i \in s} w_i \left[1 - \frac{(x_i - \bar{x}_w)(\bar{x}_w - \bar{X}_N)}{\sum_{i \in s} w_i (x_i - \bar{x}_w)^2} \right] y_i + o_p(n^{-\frac{1}{2}}) = \bar{y}_{\text{GREG}} + o_p(n^{-\frac{1}{2}}),$$

where $\bar{x}_w = \sum_{i \in s} w_i x_i$.

4 Simulation results

We study by simulation the finite sample properties of various estimators. We basically adopt the models used by Chen and Qin (1993) so that we are able to compare the cross-entropy minimization estimates with existing methods. The models we considered are as follows:

Model 1. $y = x + Zx^{\frac{1}{2}}/5$,

Model 2. $y = x + 0.05x^2 + Zx^{\frac{1}{2}}/5$,

Model 3. $y = 1.5 + x + Zx^{\frac{1}{2}}/5$,

Model 4. $y = 3 + x - 0.05x^2 + Zx^{\frac{1}{2}}/5$,

Model 5. $y = 5 + Zx^{\frac{1}{2}}/5$,

where Z in the above models is a random variable with standard normal distribution. In this simulation, x is also randomly generated and has a $\chi^2_{(6)}/2$ distribution. The variance structure is the same as in Robinson's paper. We set the finite population size $N = 1,000$ and let the sample size $n = 31$.

Note that the linear relationship between y and x gradually disappears from the Models 1 to 5. We therefore expect the ratio estimates or the estimates derived from the ratio model to get worse and worse. we only consider the estimation of $\bar{Y}$ when $\bar{X}$ is known.

Table 1 Comparing MSE's of the CEME, EMLE, RAE and RE under SRSWOR

M odel	MSE(CE)/MSE(EM)	MSE(CE)/MSE(RAE)	MSE(CE)/MSE(RE)
1	0.939	1.101	1.111
2	0.917	0.677	0.659
3	0.819	0.113	0.124
4	0.782	0.035	0.037
5	0.804	0.011	0.012

Table 2 The mean absolute errors of $\hat{Y}_{EL}$, $\hat{Y}_{CE}$ and $\bar{y}$

Model	Mabs($\bar{y}$)	Mabs($\hat{Y}_{EL}$)	Mabs($\hat{Y}_{CE}$)
$\chi^2_{(6)}/2$	0.2409	0.0596	0.0589
$\chi^2_{(10)}$	0.6529	0.2406	0.2342

The general strategy of our simulation is as follows. We use a random number generator to create a finite population of size $N = 1,000$ for each model. Then $R = 1,000$ times we draw a simple random sample without replacement of size $n = 31$ from this population, calculating the various estimates. Our simulation results are reported in the following Table 1. The numbers in the first column refer to the model being used. The columns corresponding to the ratios of $MSE(\hat{Y}_{CE})$ to $MSE(\hat{Y}_{EL})$, $MSE(\hat{Y}_{RA})$ and $MSE(\hat{Y}_R)$ .The simulation mean square errors of the four estimators were calculated as $MSE_j = \sum\limits_{k=1}^{R} \{\hat{Y}_k^{(j)} - \bar{Y}\}^2/R$, where $\hat{Y}_k^{(j)}$ is the value of $\hat{\hat{Y}}_k$ for the kth simulation run and $j = 1, 2, 3, 4,$ refer to the CEME, the EMLE, the Robinson's adjust ratio estimator (RAE) and the ratio estimator (RE), respectively. The definitions of $\hat{Y}_{RA}$ and $\hat{Y}_R$ are given by Robinson (1987).

Under model 4, and let x have $\chi^2_{(6)}/2$ distribution and $\chi^2_{(10)}$ distribution, respectively. The strategy of our simulation is similar to the above one. We calculate the mean absolute errors of $\hat{Y}_{EL}$, $\hat{Y}_{CE}$ and sample mean $\bar{y}$ under R=1,000 times simulation. The results are given in Table 2.

1. From Table 1, the cross-entropy estimates perform uniformly well. They are best in models 3, 4, and 5, when there are certain degrees of model departures from the ratio model. They are not significantly worse than any other estimates when the ratio model is approximately correct.
2. From Table 2, we can see the cross-entropy estimator is also more robust than its competitors.

Appendix: Some proofs

Proof of Theorem 2.1 It is easy to show that the function of the left side of Eq. (6) be monotone. Thus the solution λ of Eq. (6) exists and is unique. Then we have

$$\begin{aligned} 0 &= \frac{1}{n}\sum_{i\in s} u_i \exp\{\lambda u_i\} \\ &= \frac{1}{n}\sum_{i\in s} u_i[1+\lambda u_i + o(\lambda u_i)] \\ &= \frac{1}{n}\sum_{i\in s} u_i + \lambda\cdot\frac{1}{n}\sum_{i\in s} {u_i}^2 + \lambda\cdot o\left(\frac{1}{n}\sum_{i\in s} {u_i}^2\right) \\ &= \bar{u} + \lambda s_u^2 + \lambda(\bar{u})^2 + \lambda\cdot o\left(\frac{1}{n}\sum_{i\in s} {u_i}^2\right), \end{aligned} \tag{10}$$

where $s_u^2 = \frac{1}{n}\sum_{i\in s}(u_i-\bar{u})^2, \quad \bar{u} = \frac{1}{n}\sum_{i\in s} u_i$. Since $N^{-1}\sum_{i=1}^N |u_i|^3 = O_p(1)$, we have

$$p_r\{|\bar{u}| > c(n^{-1}-N^{-1})^{\frac{1}{2}}\} \le c^{-2}(n^{-1}-N^{-1})^{-1}\mathrm{var}(\bar{u}) = c^{-2}(N-1)^{-1}\sum_{i=1}^N u_i^2,$$

which implies $\bar{u} = O_p\{(n^{-1}-N^{-1})^{\frac{1}{2}}\} = O_p(n^{-\frac{1}{2}})$. From Eq. (10) and the fact

$$\left\{\frac{1}{n}\sum_{i\in s} {u_i}^2\right\}^{-1} = O_p(1),$$

we have

$$\lambda = -{s_u}^{-2}\bar{u} + o_p\{(n^{-1}-N^{-1})^{\frac{1}{2}}\} = O_p\{(n^{-1}-N^{-1})^{\frac{1}{2}}\} = O_p(n^{-\frac{1}{2}}). \tag{11}$$

From Eq. (11) and

$$\begin{aligned} \sum_{i\in s}\exp(\lambda u_i) &= \sum_{i\in s}[1+\lambda u_i + o(\lambda u_i)] \\ &= n\left[1+\lambda\cdot\frac{1}{n}\sum_{i\in s} u_i + o\left(\lambda\cdot\frac{1}{n}\sum_{i\in s} u_i\right)\right] = n[1+O_p(n^{-1})], \end{aligned}$$

it is easily to find that

$$\begin{aligned} \hat{\theta}_n &= \sum_{i\in s}\hat{p}_i g_i \\ &= \left(\sum_{i\in s} g_i\exp\{\lambda u_i\}\Big/\left(\sum_{i\in s}\exp\{\lambda u_i\}\right)\right. \\ &= \frac{1}{n}\left[1+O_p(n^{-1})\right]\left[\sum_{i\in s} g_i(1+\lambda u_i + o_p(\lambda u_i))\right] \\ &= \bar{g} - \frac{S_{gu}}{S_u^2}\bar{u} + o_p(n^{-1}) \\ &= \bar{g} - \frac{\sigma_{gu}}{\sigma_u^2}\bar{u} + o_p(n^{-1}), \end{aligned} \tag{12}$$

where $S_{gu} = n^{-1}\sum_{i\in s}(u_i-\bar{u})(g_i-\bar{g})$. This implies the desired result. □

Proof of Theorem 2.2 Suppose that λ_{-j} is the Lagrange multiplier with the jth unit removed. From

$$\sum_{i\neq j} u_i(\exp\{\lambda_{-j}u_i\} - \exp\{\lambda u_i\}) = u_j \exp\{\lambda u_j\},$$

we get

$$\lambda_{-j} - \lambda = \frac{u_j}{n\sigma_u^2}\{1 + o_p(1)\},$$

where the $o_p(1)$ is uniform over j. Hence

$$\begin{aligned}
\hat{\theta}_{-j} - \hat{\theta}_n &= \frac{\sum_{i\neq j} g_i \exp\{\lambda_{-j}u_i\}}{\sum_{i\neq j}\exp\{\lambda_{-j}u_i\}} - \frac{\sum_{i\in s} g_i \exp\{\lambda_i u_i\}}{\sum_{i\in s}\exp\{\lambda_i u_i\}} \\
&= \frac{n(\sum_{i\neq j} g_i \exp\{\lambda_{-j}u_i\}) - (n-1)(\sum_{i\in s} g_i \exp\{\lambda u_i\})}{n(n-1)} \cdot (1 + o_p(1)) \\
&= \frac{n[\sum_{i\neq j} g_i(\exp\{\lambda_{-j}u_i\} - \exp\{\lambda_i u_i\})] + \sum_{i\neq j} g_i \exp\{\lambda u_i\} - (n-1)g_j \exp\{\lambda u_j\}}{n(n-1)} \\
&\quad \cdot(1 + o_p(1)) \\
&= \left[\frac{1}{n-1}\sum_{i\neq j} g_i u_i + o_p(1)\right](\lambda_{-j} - \lambda) + \frac{1}{n(n-1)}\left[\sum_{i\neq j} g_i(1 + o_p(1))\right] \\
&\quad -\frac{1}{n} g_j(1 + o_p(1)) \\
&= -\frac{1}{n-1}[g_j(1 + o_p(1)) - \bar{g} - u_j\left\{\frac{\sigma_{gu}}{\sigma_u^2} + o_p(1)\right\}],
\end{aligned}$$

where, again, $o_p(1)$ is uniform over j, and then

$$\begin{aligned}
\sigma_{\mathrm{J}}^2 &= \left(1 - \frac{n}{N}\right)(n-1)\sum_{j\in s}(\hat{\theta}_{-j} - \hat{\theta}_n)^2 \\
&= \left(1 - \frac{n}{N}\right)(n-1)^{-1}\sum_{j\in s}\left[g_j\{1 + o_p(1)\} - \bar{g} - u_j\left\{\frac{\sigma_{gu}}{\sigma_u^2} + o_p(1)\right\}\right]^2 \\
&= \sigma_v^2 + o_p(1).
\end{aligned}$$

This implies the result. □

Proof of Theorem 3.1 From the conditions of Theorem 3.1 and

$$\begin{aligned}
0 &= \Big(\sum_{i\in s} d_i\Big)^{-1}\sum_{i\in s} d_i u_i \exp\{\lambda u_i\} \\
&= \Big(\sum_{i\in s} d_i\Big)^{-1}\sum_{i\in s} d_i u_i(1 + \lambda u_i + o(\lambda u_i)) \\
&= \sum_{i\in s} w_i u_i + \lambda\sum_{i\in s} w_i u_i^2 + o\Big(\lambda\sum_{i\in s} w_i u_i^2\Big),
\end{aligned}$$

we must have

$$\lambda = S_{wu}^{-2}\bar{u}_w + o_p(n^{-\frac{1}{2}}) = O_p(n^{-\frac{1}{2}}), \tag{13}$$

where $\bar{u}_w = \sum_{i\in s} w_i u_i, \quad S_{wu}^2 = \sum_{i\in s} w_i(u_i - \bar{u}_w)^2$. Using the Taylor expansion, we have

$$\sum_{i\in s} d_i \exp\{\lambda u_i\} = \sum_{i\in s} d_i[1 + \lambda u_i + o(\lambda u_i)] = \left(\sum_{i\in s} d_i\right)[1 + \lambda\bar{u}_w + o(\lambda\bar{u}_w)].$$

From Eq. (13) and condition (iii), we conclude that $\lambda\bar{u}_w = o_p(1)$, and thus

$$\begin{aligned}\hat{\theta}_n &= \sum_{i\in s}\hat{p}_i g_i \\ &= \Big(\sum_{i\in s} d_i \exp\{\lambda u_i\}\Big)^{-1}\Big(\sum_{i\in s} d_i g_i \exp\{\lambda u_i\}\Big) \\ &= \Big[\sum_{i\in s} d_i(1 + o_p(1))\Big]^{-1}\Big[\sum_{i\in s} d_i g_i(1 + \lambda u_i + o(\lambda u_i))\Big] \\ &= \bar{g}_w - \frac{S_{wug}}{S_{wu}^2}\bar{u}_w + o_p(n^{-\frac{1}{2}}),\end{aligned}$$

where $S_{wug} = \sum_{i\in s} w_i(u_i - \bar{u}_w)(g_i - \bar{g}_w)$. We then obtain the result of Theorem 3.1. □

Acknowledgements This paper was partially supported by the NNSF of China grant No. 10571093 and the SRFDP of China grant No. 20050055038.

References

Bero, A. K., Bilias, Y. (2002). The MM, ME, ML, EL, EF and GMM approaches to estimation: a synthesis. *Journal of Econometrics*, *107*, 51–86.

Bickel, P. J., Freedman, D. A. (1984). Asymptotic normality and the bootstrap in stratified sampling. *The Annals of Statistics*, *12*, 470–482.

Chen, J., Qin, J. (1993). Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. *Biometrika*, *80*, 107–116.

Chen, J., Sitter, R. R. (1999). A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statistica Sinica*, *9*, 384–406.

Deville, J. C., Särndal, C. E. (1992). Calibration estimators in survey sampling. *Journal of the Americam Statistical Association*, *87*, 376–382.

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

Hellerstein, J., Imbens, G. W.(1999). Imposing moment restrictions from auxiliary data by weighting. *Review of Economics and Statistics*, *81*, 1–14.

Imbens, G. W., Spady, R. H., Johnson, P. (1998). Information theoretic approaches to inference in moment condition models. *Econometrica*, *66*, 333–357.

Owen, A. B. (1990). Empirical likelihood confidence regions. *The Annals of Statistics*, *18*, 90–120.

Ullah, A. (1996). Entropy, divergence and distance measures with econometric applications. *Journal of Statistical Planing and Inference*, *49*, 137–162.