Statistica Sinica **19** (2009), 145-157

# CALIBRATION ESTIMATION USING EMPIRICAL LIKELIHOOD IN SURVEY SAMPLING

Jae Kwang Kim

*Iowa State University*

*Abstract:* Calibration estimation, which can be roughly described as a method of adjusting the original design weights to incorporate the known population totals of the auxiliary variables, has become very popular in sample surveys. The calibration weights are chosen to minimize a given distance measure while satisfying a set of constraints related to the auxiliary variable information. Under simple random sampling, Chen and Qin (1993) suggested that the calibration estimator maximizing the constrained empirical likelihood can make efficient use of the auxiliary variables. We extend the result to unequal probability sampling and propose an algorithm to implement the proposed procedure. Asymptotic properties of the proposed calibration estimator are discussed. The proposed method is extended to the stratified sampling. Results from a limited simulation study are presented.

*Key words and phrases:* Generalized regression estimator, nonparametric maximum likelihood estimator, optimal regression estimator, weighting procedure.

## 1. Introduction

In samples selected from a finite population, auxiliary variables with known population totals are often observed. The known population totals usually come from external sources such as administrative data or a census. Calibration estimation, which can be roughly described as a method of adjusting the original design weights to incorporate the known population totals of the auxiliary variables, has become very popular in sample surveys. Generally speaking, the calibration procedure chooses the adjusted weights that minimize a distance between the original weights and the adjusted weights, while satisfying a set of constraints related to the auxiliary variable information. Fuller (2002) provides a comprehensive overview of the calibration procedure in sample surveys.

From a purely mathematical point of view, the calibration estimation problem is a standard optimization problem with constraints and, given the constraints, the choice of the objective function determines the properties of the resulting estimator. The classical regression estimator described in Cochran (1977) uses a Euclidian distance function. Deville and Särndal (1992) gave conditions for the distance functions to produce calibration estimators that are asymptotically equivalent to the regression estimator.

In addition to the above interpretation of minimizing a distance function, the calibration estimator can sometimes be viewed as a maximum likelihood estimator. Anderson (1957) derived the regression estimator as a maximum likelihood estimator under the bivariate normal distribution assumption. Hartley and Rao (1968) used a multinomial distribution for distinct sample values and proposed a scale-load estimator that can be obtained as a constrained maximum likelihood estimator. The empirical likelihood, so named by Owen (1988), is essentially the likelihood of the multinomial distribution used in Hartley and Rao (1968), where the parameters are the point masses assigned to the distinct sample values. Under simple random sampling, Chen and Qin (1993) proposed a calibration estimator that maximizes the empirical likelihood with constraints. Chen and Sitter (1999) extended the method to unequal probability designs using a pseudo empirical likelihood function. In this paper, we propose a new type of empirical likelihood calibration estimator that preserves the maximum likelihood interpretation under Poisson sampling. The hope is that the resulting estimator is still efficient under other unequal probability sampling. The objective function we consider is different from that of Chen and Sitter (1999) and thus the two estimators have different asymptotic properties.

In Section 2, the basic setup is introduced and the proposed method is described. In Section 3, asymptotic properties of the proposed estimator are discussed. The proposed method is extended to the stratified sampling in Section 4. In Section 5, results from a simulation study are presented.

## 2. Empirical Likelihood Calibration Equation

We begin by introducing the notion of empirical likelihood in a simple setup. Let $y_1, \ldots, y_n$ be the outcomes of independently and identically distributed (IID) random variables from a continuous distribution function $F_0 \in \mathcal{F}$. We consider a class $\mathcal{F}_1 \subset \mathcal{F}$ of distribution functions that have support on $\{y_1, \ldots, y_n\}$. Thus, the elements in $\mathcal{F}_1$ can be written as

$$F_w(x) = \sum_{i=1}^{n} w_i I(y_i \leq x) \tag{2.1}$$

with $\sum_{i=1}^{n} w_i = 1$ and $w_i \geq 0$, where $I(y_i \leq x)$ takes the value one if $y_i \leq x$ and takes the value zero otherwise. The parameter $w_i$ is the amount of point mass that unit $y_i$ represents in the population. That is, $w_i = F_0(y_i) - F_0(y_i-)$, where $F_0$ is the true distribution function. Note that $F_w(y)$ is a distribution function, not an estimator, indexed by the set of parameters $w_1, \ldots, w_n$. For any parameter of the form $\theta = \theta(F_0)$, the estimator $\hat{F}$ of $F_0$ can be used to estimate $\theta$ by $\hat{\theta} = \theta(\hat{F})$. For a parameter $\theta$ linear in $y$ in the population, the estimator $\hat{\theta}$

using the class of distributions (2.1) leads to a linear estimator that is linear in $y$ in the sample. Linear estimation is very popular in sample surveys because it provides internal consistency between estimators for several items.

The empirical distribution function, defined for $w_i = n^{-1}$ in (2.1), given no ties, is the nonparametric maximum likelihood estimator (NPMLE) of $F_0$, since it maximizes

$$L(w) = \prod_{i=1}^{n} w_i \tag{2.2}$$

over all $w_i$'s satisfying $\sum_{i=1}^{n} w_i = 1$ and $w_i \geq 0$. Note that if the $w_i$ are known functions of a fixed number of unknown constants then (2.2) is the usual parametric likelihood function. For any parameter of the form $\theta = \theta(F)$, the NPMLE $\hat{F}$ of $F_0$ can be used to compute the NPMLE of $\theta$ by $\hat{\theta} = \theta(\hat{F})$.

If we observe the auxiliary variable $x_i$ in the sample and the population mean of $x_i$ is known, denoted by $\mu_x$, the additional information of $\mu_x$ can be used to construct a constrained NPMLE of $F_0$. Chen and Qin (1993) proposed computing the constrained NPMLE of $F_0$ by solving

$$\text{maximize} \sum_{i=1}^{n} \log(w_i) \tag{2.3}$$

subject to

$$\sum_{i=1}^{n} w_i (1, x_i) = (1, \mu_x) \text{ and } w_i \geq 0, \forall i. \tag{2.4}$$

The constrained NPMLE for $\theta = \theta(F_0)$ can be computed from the NPMLE of $\theta$ by $\hat{\theta} = \theta\left(\hat{F}^*\right)$, where $\hat{F}^*$ is the cumulative distribution function using the solution $w_i^*$ in (2.3) and (2.4).

We now consider an extension of the constrained NPMLE to samples selected from a finite population with unequal selection probabilities. Since direct computation of the empirical likelihood function involves higher order inclusion probabilities, we consider an approximation by Poisson sampling. Let $Y_1, \ldots, Y_N$ be the vector of realized values of the finite population with the cumulative distribution function $F_0 = N^{-1} \sum_{i=1}^{N} I(Y_i \leq x)$. Since we have assumed $F_0 \in \mathcal{F}_1$, we can write

$$F_0(x) = \sum_{i=1}^{n} w_i I(y_i \leq x) \tag{2.5}$$

for some $w_i$'s. Assume that the sample is a result of $N$ independent Bernoulli trials where $\pi_i = \pi(Y_i)$ is the probability of selecting unit $i$. If we use $y$ to denote the sample value and use $Y$ to denote the population value, the sample distribution function under Poisson sampling can be written

$$\begin{aligned}\Pr(y \le x) &= \Pr[Y \le x \mid U \le \pi(Y)] \\ &= \frac{\Pr[Y \le x, U \le \pi(Y)]}{\Pr[U \le \pi(Y)]} \\ &= \frac{\int_{-\infty}^{x}\int_{0}^{\pi(Y)} du dF_0}{\int_{-\infty}^{\infty}\int_{0}^{\pi(Y)} du dF_0},\end{aligned}$$

where $U$ is a random variable whose distribution is $U(0,1)$. By (2.5), the above probability can be written

$$\Pr(y \le x) = \frac{\sum_{\{i;\ y_i \le x\}} \pi_i w_i}{\sum_{j=1}^{n} \pi_j w_j}.$$

Thus, the empirical likelihood under Poisson sampling can be written

$$L(w) = \prod_{i=1}^{n} \left( \frac{\pi_i w_i}{\sum_{j=1}^{n} \pi_j w_j} \right), \tag{2.6}$$

with $\sum_{i=1}^{n} w_i = 1$, $w_i \ge 0$, and, without loss of generality, the first $n$ elements are selected. The maximum likelihood estimator of $w_i$ using the empirical likelihood (2.6) is

$$w_i^* = \frac{\pi_i^{-1}}{\sum_{j=1}^{n} \pi_j^{-1}}, \tag{2.7}$$

which reduces to the Hájek estimator of the population mean. The empirical likelihood function in (2.6) can be found in length-biased sampling, where $\pi_i \propto Y_i$. See, for example, Vardi (1985), Qin (1993) and Kong et al. (2003). When the unequal probability sampling design is well approximated by a Poisson sampling, the empirical likelihood function (2.6) can be a good approximation to the actual likelihood function.

Using the likelihood function (2.6), the empirical likelihood calibration estimator can be derived as a constrained NPMLE for the distribution function of the finite population. The constrained maximization problem can be formulated as maximizing (2.6) subject to the constraints in (2.4). The objective function to be minimized is, by the Lagrange multiplier method,

$$Q(w) = \sum_{i=1}^{n} \log(\pi_i w_i) - n \log\left(\sum_{i=1}^{n} \pi_i w_i\right) - \lambda_1\left(\sum_{i=1}^{n} w_i - 1\right) - n\lambda_2\left(\sum_{i=1}^{n} w_i x_i - \mu_x\right).$$

Setting the partial derivative of $Q$ with respect to $w_i$ equal to zero and using $\sum_{i=1}^n w_i\left(\partial Q/\partial w_i\right)=0$, the solution satisfies

$$w_i^* = \frac{1}{n\Big(\frac{\pi_i}{\sum_j w_j^*\pi_j}\Big)+n\lambda_2\left(x_i-\mu_x\right)}. \tag{2.8}$$

Therefore, the constrained NPMLE of $w_i$ can be written

$$w_i^* = \frac{1}{n}\times\frac{1}{\hat{\lambda}_1\tilde{\pi}_i+\hat{\lambda}_2\left(x_i-\mu_x\right)}, \tag{2.9}$$

where $\tilde{\pi}_i=(\hat{N}/n)\pi_i$ with $\hat{N}=\sum_{i=1}^n \pi_i^{-1}$, and the $\lambda_i$ $(i=1,2)$ are the solutions to

$$\sum_{i=1}^n w_i^*\left(1,x_i\right)=\left(1,\mu_x\right) \tag{2.10}$$

with $w_i^*>0$ for all $i=1,\ldots,n$. Note that $w_i^*$ defined at (2.9) with (2.10) satisfy (2.8). A modified Newton-Raphson method can be used to solve the nonlinear equations (2.10). See, for example, Chen, Sitter and Wu (2002).

Chen and Sitter (1999) also considered unequal probability sampling and proposed the pseudo empirical likelihood estimator. Instead of maximizing (2.6), they proposed maximizing

$$L\left(w\right)=\sum_{i=1}^n\frac{1}{\pi_i}\log\left(w_i\right), \tag{2.11}$$

subject to the same constraints (2.4). The resulting pseudo empirical maximum likelihood estimator (PEMLE) for the mean of $y$ is $\bar{y}_{PEMLE}=\sum_{i=1}^n w_i^* y_i$ where

$$w_i^*=\frac{1}{\pi_i\left(\hat{\lambda}_1+\hat{\lambda}_2 x_i\right)} \tag{2.12}$$

and the $\hat{\lambda}_k$ $(k=1,2)$ satisfy (2.4). When the sampling mechanism is well approximated by a Poisson sampling, we expect that our method is more efficient in large samples. Efficiency will be investigated further in the next section.

## 3. Asymptotic Properties

We now study the asymptotic properties of the calibration NPMLE estimator of the population mean. To discuss the asymptotic properties of the empirical likelihood estimator, assume a sequence of finite populations with finite fourth moments as defined in Isaki and Fuller (1982).

Assume the sampling mechanism satisfies

$$K_1 < \max_i \left\{n^{-1} N \pi_i\right\} < K_2 \tag{3.1}$$

for some positive constants $K_1$ and $K_2$. Define $u_i = x_i - \mu_x$ and assume that

$$\max_i |u_i| = o_p\left(n^{\frac{1}{2}}\right), \tag{3.2}$$

$$\frac{\sum_{i=1}^n \pi_i^{-1} u_i}{\sum_{i=1}^n \pi_i^{-1} u_i^2} = O_p\left(n^{-\frac{1}{2}}\right). \tag{3.3}$$

Under the assumptions (3.1)-(3.3), Chen and Sitter (1999) proved that their pseudo empirical likelihood estimator is asymptotically equivalent to the generalized regression (GREG) estimator

$$\bar{y}_{GREG} = \bar{y}_\pi + (\mu_x - \bar{x}_\pi)\hat{B}, \tag{3.4}$$

where

$$(\bar{x}_\pi, \bar{y}_\pi) = \left(\sum_{i=1}^n \pi_i^{-1}\right)^{-1} \sum_{i=1}^n \pi_i^{-1}(x_i, y_i),$$

$$\hat{B} = \frac{\sum_{i=1}^n \pi_i^{-1}(x_i - \bar{x}_\pi)(y_i - \bar{y}_\pi)}{\sum_{i=1}^n \pi_i^{-1}(x_i - \bar{x}_\pi)^2}.$$

The following theorem states some asymptotic properties of the calibration NPMLE using the weights in (2.9) with (2.10).

**Theorem 1.** *Under the assumptions* (3.1)−(3.3), *the NPMLE of the mean of $y$ is asymptotically equivalent to*

$$\bar{y}_{opt} = \bar{y}_\pi + (\mu_x - \bar{x}_\pi)\hat{B}^*, \tag{3.5}$$

*where* $\hat{B}^* = \sum_{i=1}^n \pi_i^{-2}(x_i - \bar{x}_\pi)(y_i - \bar{y}_\pi) / \sum_{i=1}^n \pi_i^{-2}(x_i - \bar{x}_\pi)^2$, *and* $(\bar{x}_\pi, \bar{y}_\pi)$ *is defined after* (3.4).

The proof of the Theorem is given in Appendix A.

Note that the NPMLE is motivated by Poisson sampling, but the result in Theorem 1 does not require it. Since $\hat{B}^* = O_p(1)$, consistency of the NPMLE follows directly. Under Poisson sampling,

$$\hat{C}(\bar{x}_\pi, \bar{y}_\pi) = \hat{N}^{-2} \sum_{i=1}^n \left(\pi_i^{-2} - \pi_i^{-1}\right)(x_i - \bar{x}_\pi)(y_i - \bar{y}_\pi),$$

$$\hat{V}(\bar{x}_\pi) = \hat{N}^{-2} \sum_{i=1}^n \left(\pi_i^{-2} - \pi_i^{-1}\right)(x_i - \bar{x}_\pi)^2,$$

with $\hat{N} = \sum_{i=1}^{n} \pi_i^{-1}$, are consistent estimators of $Cov\,(\bar{x}_\pi, \bar{y}_\pi)$ and $Var\,(\bar{x}_\pi)$, respectively. Thus, if the Poisson sampling design is such that

$$\max_i \ \pi_i = o\,(1)\,, \tag{3.6}$$

the $\hat{B}^*$ in (3.5) estimates $[Var\,(\bar{x}_\pi)]^{-1}\, Cov\,(\bar{x}_\pi, \bar{y}_\pi)$. Thus, under the Poisson sampling with (3.6), the proposed NPMLE is close to the optimal estimator discussed by Rao (1994). The optimal estimator minimizes the asymptotic design variance among the class of asymptotically unbiased estimators that are linear in $(\bar{x}_\pi, \bar{y}_\pi)$. The idea of using $\pi_i^{-2}$ to compute the regression coefficient also appears in Isaki and Fuller (1982)

## 4. Extension to Stratified Sampling

The proposed NPMLE method can be extended to stratified sampling with unequal probability of selection in each stratum. Let the finite population of $N$ units be partitioned into $H$ strata with known stratum sizes $N_1, \ldots, N_H$. In stratum $h$, we observe $y_{hi}$ with the probability of selection $\pi_{hi}$, for $i = 1, \ldots, n_h$. Here, we assume that the first $n_h$ elements are sampled in each stratum. In addition to $y_{hi}$ we also observe $x_{hi}$, and only the population mean $\mu_x$ of $x_{hi}$ is known.

Let $w_{hi}$ be the proportion that unit $y_{hi}$ represents in the population in stratum $h$. Thus, the NPMLE can be formulated as maximizing

$$L\,(w) = \prod_{h=1}^{H} \prod_{i=1}^{n_h} \left( \frac{\pi_{hi} w_{hi}}{\sum_{j=1}^{n_h} \pi_{hj} w_{hj}} \right), \tag{4.1}$$

subject to

$$\sum_{i=1}^{n_h} w_{hi} = 1, \quad h = 1, \ldots, H, \tag{4.2}$$

$$\sum_{h=1}^{H} W_h \sum_{i=1}^{n_h} w_{hi} x_{hi} = \mu_x, \tag{4.3}$$

and $w_{hi} \geq 0$, where $W_h = N_h/N$. Using the Lagrange multiplier method again, the solution $w_{hi}^*$ can be expressed as

$$w_{hi}^* = \frac{1}{n_h} \times \frac{1}{\hat{\lambda}_h \tilde{\pi}_{hi} + \hat{\lambda}_{H+1} m_h\,(x_{hi} - \tilde{x}_h)}, \tag{4.4}$$

where $\tilde{\pi}_{hi} = (\hat{N}_h/n_h)\pi_{hi}$, $\hat{N}_h = \sum_{i=1}^{n_h} \pi_{hi}^{-1}$, $m_h = W_h(n/n_h)$, $\tilde{x}_h = \sum_{i=1}^{n_h} w_{hi}^* x_{hi}$ and $\hat{\lambda}_h\,(h = 1, \ldots, H, H+1)$ are the solution to (4.2) and (4.3). To compute

$\hat{\lambda}_h$'s, as in Chen and Sitter (1999), we first express (4.3) as a single function of $\lambda_{H+1}$ and obtain $\hat{\lambda}_{H+1}$ first by the bisection method. The $\hat{\lambda}_h$ $(h = 1, \ldots, H)$ are computed directly by (4.2). The following theorem provides some asymptotic properties of the NPMLE of the mean of $y$ in stratified sampling.

**Theorem 2.** *For the sequence of stratified populations and samples described in* Chen and Sitter (1999) *with $H$ fixed, the NPMLE of the mean of $y$ in stratified sampling is asymptotically equivalent to*

$$\bar{y}_{opt} = \bar{y}_\pi + (\mu_x - \bar{x}_\pi)\,\hat{B}^*_{st}, \tag{4.5}$$

*where* $(\bar{x}_\pi, \bar{y}_\pi) = \sum_{h=1}^H W_h\,(\bar{x}_h, \bar{y}_h)$, $(\bar{x}_h, \bar{y}_h) = \sum_{i=1}^{n_h} d_{hi}\,(x_{hi}, y_{hi})$, $d_{hi} = \pi_{hi}^{-1} / \sum_{j=1}^{n_h} \pi_{hj}^{-1}$, *and*

$$\hat{B}^*_{st} = \frac{\sum_{h=1}^H W_h^2 \sum_{i=1}^{n_h} d_{hi}^2\,(x_{hi} - \bar{x}_h)\,(y_{hi} - \bar{y}_h)}{\sum_{h=1}^H W_h^2 \sum_{i=1}^{n_h} d_{hi}^2\,(x_{hi} - \bar{x}_h)\,(x_{hi} - \bar{x}_h)}.$$

The proof of the Theorem is given in Appendix B.

If stratified random sampling is used, $d_{hi} = n_h^{-1}$ and

$$\hat{B}^*_{st} = \frac{\sum_{h=1}^H W_h^2 n_h^{-2} \sum_{i=1}^{n_h} (x_{hi} - \bar{x}_h)\,(y_{hi} - \bar{y}_h)}{\sum_{h=1}^H W_h^2 n_h^{-2} \sum_{i=1}^{n_h} (x_{hi} - \bar{x}_h)\,(x_{hi} - \bar{x}_h)},$$

which is asymptotically equivalent to the estimator of Zhong and Rao (2000), and thus is asymptotically equivalent to the optimal estimator in stratified random sampling. A more comprehensive treatment of the NPMLE under stratified sampling is a topic of future research.

## 5. Simulation Studies

To study the properties of the proposed calibration estimator, we performed a limited simulation study. In the simulation study, four artificial finite populations of size $N = 10,000$ were generated:

(A) $z_i \sim \chi^2(2)$, $x_i = a_i + 0.5z_i + 2$, $y_i = 1 + \sqrt{0.5}\,(x_i - 3) + e_i$;

(B) $(x_i, z_i)$ are the same as in population [A] and $y_i = (x_i - 3)^2 + e_i$;

(C) $z_i \sim \chi^2(2) + 2$, $x_i = a_i + 0.5z_i + 1$, $y_i = 1 + \sqrt{0.5}\,(x_i - 3) + e_i$;

(D) $(x_i, z_i)$ are the same as in population [C] and $y_i = (x_i - 3)^2 + e_i$.

In the four populations $a_i \sim N(0,1)$, independent of $z_i$, and $e_i \sim N(0,1)$, independent of $(u_i, a_i, z_i)$. Thus, the population values of $(x_i, y_i)$ in populations A and B are essentially the same as those in populations C and D, respectively.

From each of the finite populations generated above, $M = 5,000$ Monte Carlo samples of size $n$ were generated by probability proportional to size (PPS) sampling with replacement. The finite populations were fixed in the Monte Carlo sampling. In the PPS sampling, we allowed for duplication of the population elements in the sample and the probability of selecting a single element $p_i$ was proportional to $z_i$; two sample sizes, $n = 200$ and $n = 500$, were used. Thus, the sampling weights, $w_i = N^{-1}n^{-1}p_i^{-1}$ where $p_i = z_i / \left(\sum_{i=1}^{N} z_i\right)$, in populations A and B are more extreme than those in populations C and D because the values of $z_i$ in populations A and B are generally smaller than those in populations C and D. We assumed that the population mean of $x_i$ was known and was used for the calibration. From each sample, four estimators of the population mean of $y$ were computed. The estimators are the Hansen-Hurwitz (HH) estimator for the PPS sampling, the GREG estimator defined in (3.4), the pseudo empirical likelihood estimator (PEMLE) of Chen and Sitter (1999) defined in (2.12), and the proposed NPMLE defined in (2.9).

Monte Carlo biases and Monte Carlo mean squared errors were computed for the four point estimators. Table 1 reports the simulation results of the four point estimators. From the results in Table 1, we have the following conclusions.

1. There are modest biases in the point estimators. The magnitude of the biases is much smaller than the standard error, and the bias is smaller for $n = 500$ than for $n = 200$. Because the bias is of order $O\left(n^{-1}\right)$, the bias can be safely ignored in the asymptotic sense.
2. The HH estimators in population A and B have bigger variances than those in population C and D. Since the reciprocal of the $z$-variable is highly variable in population A and B, the resulting design weights for the HH estimator are also highly variable. Since the design weights are independent of the $y$-variable, the extreme weights increase the variances of the resulting HH estimators.
3. In population C, the three calibration estimators show similar performances because the weights are relatively homogeneous. The ratio of the variance of the calibration estimator to the variance of the HH estimator is about 0.5, which is consistent with the theory because the population correlation between $x$ and $y$ is $\sqrt{0.5}$.
4. In populations A and B, the NPMLE shows better performance than the other calibration estimators. Note that the two empirical likelihood estimators can be written

$$\hat{\theta}_{PEMLE} = \sum_{i=1}^{n} \frac{d_i y_i}{\lambda_1 + \lambda_2 \left(x_i - \mu_x\right)},$$

$$\hat{\theta}_{NPMLE} = \sum_{i=1}^{n} \frac{d_i y_i}{\lambda_1 + \lambda_2 d_i \left(x_i - \mu_x\right)}.$$

Table 1. Monte Carlo Biases and Monte Carlo Mean squared errors of the point estimators, based on 5,000 Monte Carlo samples.

| n | Pop'n | Estimator | Bias | MSE |
|---|---|---|---|---|
| 200 | A | HH | 0.00 | 0.0522 |
| | | GREG | 0.00 | 0.0183 |
| | | PEMLE | 0.00 | 0.0188 |
| | | NPMLE | 0.00 | 0.0179 |
| | B | HH | 0.01 | 0.2877 |
| | | GREG | -0.08 | 0.1872 |
| | | PEMLE | 0.05 | 0.3860 |
| | | NPMLE | -0.01 | 0.0972 |
| | C | HH | 0.00 | 0.00926 |
| | | GREG | 0.00 | 0.00612 |
| | | PEMLE | 0.00 | 0.00612 |
| | | NPMLE | 0.00 | 0.00615 |
| | D | HH | 0.00 | 0.0414 |
| | | GREG | -0.02 | 0.0537 |
| | | PEMLE | 0.01 | 0.0546 |
| | | NPMLE | 0.00 | 0.0459 |
| 500 | A | HH | 0.00 | 0.01905 |
| | | GREG | 0.00 | 0.00840 |
| | | PEMLE | 0.00 | 0.00850 |
| | | NPMLE | 0.00 | 0.00804 |
| | B | HH | 0.00 | 0.4911 |
| | | GREG | -0.04 | 0.0873 |
| | | PEMLE | 0.03 | 0.1718 |
| | | NPMLE | -0.01 | 0.0427 |
| | C | HH | 0.00 | 0.00359 |
| | | GREG | 0.00 | 0.00236 |
| | | PEMLE | 0.00 | 0.00236 |
| | | NPMLE | 0.00 | 0.00236 |
| | D | HH | 0.00 | 0.0161 |
| | | GREG | -0.01 | 0.0211 |
| | | PEMLE | 0.01 | 0.0212 |
| | | NPMLE | 0.00 | 0.0179 |

Thus, the PEMLE will be efficient if $d_i y_i \propto x_i - \mu_x$, while the NPMLE will be efficient if $d_i y_i \propto d_i (x_i - \mu_x)$, or $y_i \propto x_i$. If the design weights $d_i$ are highly variable, the PEMLE can be inefficient. Therefore, the NPMLE is less sensitive to extreme design weights.

5. In population D, where design weights are relatively homogeneous and the linear relationship between $y$ and $x$ does not hold, the calibration estimators

do not improve the efficiency of the HH estimator. In population B, the NPMLE is more efficient than the HH estimator because the efficiency of the HH estimator is mitigated by extreme weights.

**Acknowledgement**

The research was performed when the author was working at Yonsei university in Korea. This work was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD) (KRF-2005-041-C00086). The author thanks Tae-Hoon Lee, anonymous referees, and an associate editor for very helpful comments.

## Appendix

### A. Proof of Theorem 1

Let $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)'$ and define

$$\begin{aligned}\mathbf{U}(\boldsymbol{\lambda}) &= [U_1(\boldsymbol{\lambda}), U_2(\boldsymbol{\lambda})]' \\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\lambda_1 \tilde{\pi}_i + \lambda_2 (x_i - \mu_x)} (1, x_i)' - (1, \mu_x)',\end{aligned}$$

where $\tilde{\pi}_i = (\hat{N}/n)\pi_i$. Then $\hat{\boldsymbol{\lambda}} = \left(\hat{\lambda}_1, \hat{\lambda}_2\right)'$ is the solution to $U(\boldsymbol{\lambda}) = \mathbf{0}$. Using the argument of Owen (1990, pp.100-101), it can be shown that $\hat{\boldsymbol{\lambda}} = \boldsymbol{\lambda}_0 + O_p\left(n^{-1/2}\right)$, where $\boldsymbol{\lambda}_0 = (1, 0)'$ is the solution to $E[\mathbf{U}(\boldsymbol{\lambda})] = \mathbf{0}$. If

$$\bar{y}(\boldsymbol{\lambda}) = \frac{1}{n}\sum_{i=1}^{n} \frac{y_i}{\lambda_1 \tilde{\pi}_i + \lambda_2 (x_i - \mu_x)},$$

then $\bar{y}_{NPMLE} = \bar{y}(\hat{\boldsymbol{\lambda}})$. Taking a Taylor expansion of $\bar{y}(\hat{\boldsymbol{\lambda}})$ around $\boldsymbol{\lambda}_0$ leads to

$$\bar{y}_{NPMLE} = \bar{y}(\hat{\boldsymbol{\lambda}}) = \bar{y}(\boldsymbol{\lambda}_0) + \left(\frac{\partial \bar{y}}{\partial \boldsymbol{\lambda}}(\boldsymbol{\lambda}_0)\right)' \left(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0\right) + O_p\left(n^{-1}\right). \tag{A.1}$$

Taking a Taylor expansion of $\mathbf{U}\left(\hat{\boldsymbol{\lambda}}\right)$ around $\boldsymbol{\lambda}_0$ leads to

$$\mathbf{0} = \mathbf{U}(\hat{\boldsymbol{\lambda}}) = \mathbf{U}(\boldsymbol{\lambda}_0) + \left(\frac{\partial \mathbf{U}}{\partial \boldsymbol{\lambda}}(\boldsymbol{\lambda}_0)\right) \left(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0\right) + O_p\left(n^{-1}\right). \tag{A.2}$$

Inserting (A.2) into (A.1), we have

$$\bar{y}_{NPMLE} = \bar{y}(\boldsymbol{\lambda}_0) - \left[\frac{\partial \bar{y}}{\partial \boldsymbol{\lambda}}(\boldsymbol{\lambda}_0)\right]' \left[\frac{\partial \mathbf{U}}{\partial \boldsymbol{\lambda}}(\boldsymbol{\lambda}_0)\right]^{-1} \mathbf{U}(\boldsymbol{\lambda}_0) + O_p\left(n^{-1}\right). \tag{A.3}$$

Using $\bar{y}(\boldsymbol{\lambda}_0) = \hat{N}^{-1} \sum_{i=1}^{n} \pi_i^{-1} y_i = \bar{y}_\pi$,

$$\frac{\partial \bar{y}}{\partial \boldsymbol{\lambda}}(\boldsymbol{\lambda}_0) = -\frac{1}{n}\left[\sum_{i=1}^{n} \frac{y_i}{\tilde{\pi}_i}, \sum_{i=1}^{n} \frac{(x_i - \mu_x) y_i}{\tilde{\pi}_i^2}\right]',$$

$$\frac{\partial \mathbf{U}}{\partial \boldsymbol{\lambda}}(\boldsymbol{\lambda}_0) = -\frac{1}{n}\begin{bmatrix} \sum_{i=1}^{n} \tilde{\pi}_i^{-1} & \sum_{i=1}^{n} \tilde{\pi}_i^{-2}(x_i - \mu_x) \\ \sum_{i=1}^{n} \tilde{\pi}_i^{-1} x_i & \sum_{i=1}^{n} \tilde{\pi}_i^{-2} x_i (x_i - \mu_x) \end{bmatrix},$$

and

$$\mathbf{U}(\boldsymbol{\lambda}_0) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\tilde{\pi}_i}(1, x_i)' - (1, \mu_x)' = (0, \bar{x}_\pi - \mu_x)',$$

(A.3) reduces to

$$\bar{y}_{NPMLE} = \bar{y}_\pi + \frac{\sum_{i=1}^{n} \pi_i^{-2}(x_i - \mu_x)(y_i - \bar{y}_\pi)}{\sum_{i=1}^{n} \pi_i^{-2}(x_i - \mu_x)(x_i - \bar{x}_\pi)}(\mu_x - \bar{x}_\pi) + O_p\left(n^{-1}\right). \quad \text{(A.4)}$$

Since $\bar{x}_\pi - \mu_x = o_p(1)$, the right side of (A.4) is asymptotically equivalent to (3.5).

## B. Assumptions and Proof of Theorem 2

Let $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_H, \lambda_{H+1})'$ and $\mathbf{U}(\boldsymbol{\lambda}) = [U_1(\boldsymbol{\lambda}), \ldots, U_H(\boldsymbol{\lambda}), U_{H+1}(\boldsymbol{\lambda})]'$, where

$$U_h(\boldsymbol{\lambda}) = \frac{1}{n_h}\sum_{i=1}^{n_h} \frac{1}{\lambda_h \tilde{\pi}_{hi} + \lambda_{H+1} m_h (x_{hi} - \tilde{x}_h)} - 1, \quad h = 1, \ldots, H,$$

$$U_{H+1}(\boldsymbol{\lambda}) = \sum_{h=1}^{H} \frac{W_h}{n_h} \sum_{i=1}^{n_h} \frac{x_{hi} - \mu_x}{\lambda_h \tilde{\pi}_{hi} + \lambda_{H+1} m_h (x_{hi} - \tilde{x}_h)}.$$

Then, $\hat{\boldsymbol{\lambda}}$ be the solution to $U(\boldsymbol{\lambda}) = \mathbf{0}$. Under the regularity conditions,

$$\hat{\boldsymbol{\lambda}} = \boldsymbol{\lambda}_0 + O_p\left(n^{-\frac{1}{2}}\right),$$

where $\boldsymbol{\lambda}_0 = (1, 1, \ldots, 1, 0)'$ is the solution to $E[\mathbf{U}(\boldsymbol{\lambda})] = \mathbf{0}$. If

$$\bar{y}(\boldsymbol{\lambda}) = \sum_{h=1}^{H} \frac{W_h}{n_h} \sum_{i=1}^{n_h} \frac{y_{hi}}{\lambda_h \tilde{\pi}_{hi} + \lambda_{H+1} m_h (x_{hi} - \tilde{x}_h)},$$

then $\bar{y}_{NPMLE} = \bar{y}(\hat{\boldsymbol{\lambda}})$. Using (A.3), (4.5) follows from some matrix algebra.

## References

Anderson, T. W. (1957). Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. *J. Amer. Statist. Assoc.* **52**, 200-203.

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 estimator for surveys. *Biometrika* **89** 230-237.

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

Cochran, W. G. (1977). *Sampling Techniques.* 3rd edition. Wiley, New York.

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

Isaki, C. and Fuller, W. A. (1982). Survey design under the regression superpopulation model. *J. Amer. Statist. Assoc.* **77**, 89-96.

Hartley, H. O. and Rao, J. N. K. (1968). A new estimation theory for sample surveys. *Biometrika* **55**, 547-555.

Fuller, W. A. (2002). Regression estimation for sample surveys. *Survey Methodology* **28**, 5-23.

Kong, A., McCullagh, P., Meng, X.-L., Nicolae, D., and Tan, Z. (2003). A theory of statistical models for Monte Carlo integration. *J. Roy. Statist. Soc. Ser. B* **65**, 585-618.

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

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

Qin, J. (1993). Empirical likelihood in biased sample problems. *Ann. Statist.* **21**, 1182-1196.

Rao, J. N. K. (1994). Estimating totals and distribution functions using auxiliary information at the estimation stage. *J. Official Statist.* **10**, 153-165.

Vardi, Y. (1985). Empirical distributions in selections bias models. *Ann. Statist.* **13**, 178-203.

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, Iowa State University, Ames, IA 50011, U.S.A.

E-mail: kim00020@yahoo.com

(Received May 2006; accepted September 2007)