第27卷第1期 数 学 研 究 与 评 论 Vol.27, No.1
2007年2月 JOURNAL OF MATHEMATICAL RESEARCH AND EXPOSITION Feb., 2007

Article ID: 1000-341X(2007)01-0087-11 **Document code**: A

A Model-Calibration Information-Theoretic Approach to Using Complete Auxiliary Information

WU Chang-chun[1,2], *ZHANG Run-chu*[2]

(1. School of Mathematics and Information Science, Jiaxing University, Zhejiang 314001, China;
2. LPMC and School of Mathematical Sciences, Nankai University, Tianjin 300071, China)
(E-mail: wuchangch@163.com)

Abstract: We propose a model-calibrated K-L relative entropy minimization (MKLEM) approach to using complete auxiliary information from survey data. Our estimator is asymptotically equivalent to a model-calibration (MC) estimator in Wu and Sitter (2001) in the case of estimating the finite population mean. One attractive advantage of our MKLEM approach is the intrinsic properties of the resulting weights: $\hat{p}_i > 0$ and $\sum_{i\in s}\hat{p}_i = 1$, which make this approach generally applicable to the estimation of distribution functions and quantiles. The resulting estimator $\hat{F}_{MKL}(y)$ is asymptotically equivalent to a generalized regression estimator and itself a distribution function.

Key words: model-calibration; complete auxiliary information; K-L relative entropy; generalized regression estimator; empirical likelihood.
MSC(2000): 62D05; 62G05; 62G09; 62G20
CLC number: O212.2; O212.7

1. Introduction

In survey sampling, auxiliary information on the finite population is regularly used to increase the precision of estimators. In the simplest settings, ratio and regression estimators incorporate known finite population means of auxiliary variables. For more general situations, there have been three main methods proposed in literatures which can be categorized as model-assisted approaches: generalized regression estimators (GREG) (Cassel, Särndal and Wretman (1976) and Särndal (1980)), calibration estimator (Deville and Särndal (1992)), empirical likelihood methods (Chen and Qin (1993), Chen and Sitter (1999) and Zhong and Rao (2000)). All of these methods have only been discussed in the context of a linear regression working model and essentially incorporate the auxiliary variables through their known population means even when the auxiliary variables are known for every unit in the population.

Suppose that the finite population consists of N identifiable units. Associated with the ith unit are the study variable, y_i, and a vector of auxiliary variables, x_i. The values of $x_1, x_2, \cdots, x_N$ are known for the entire population (i.e., complete), but y_i is known only if the ith unit is selected in the sample. One of the fundamental questions is how to effectively use the complete auxiliary information at the estimation stage. More recently, a unified model-assisted framework has been

Received date: 2004-09-21; **Accepted date**: 2006-07-02
Foundation item: the National Natural Science Foundation of China (10571093); the Scientific Research Fund of Zhejiang Provincial Education Department (20061599).

attempted to use a model-calibration technique proposed by Wu and Sitter (2001). The proposed model-calibration estimator can handle any linear or nonlinear working models and reduces to the conventional calibration estimators of Deville and Särndal (1992) or the generalized regression estimators in the linear model case. They also discussed estimation of the finite population distribution functions and showed that their approaches for the mean case also provide a unified framework for the estimation of distribution function through the fitted values of the indicator variable $I_{[y\leq t]}$. However, the proposed model-calibration estimator $\hat{F}_{MC}(t)$ may take values out of the range $[0,1]$ and it is not always a monotone function. They then extend the pseudo empirical likelihood method to this situation. However, there are good reasons to explore alternatives to pseudo empirical likelihood estimator. From the definitions (Chen and Qin (1993), Chen and Sitter (1999)), we know that the defined empirical likelihood function under simple random sampling 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 iid sample from some superpopulation. Although empirical likelihood approach has desirable large sample properties, it may not be the best method. We argue that the K-L relative entropy minimization (KLEM) estimator be more appealing than the empirical maximum likelihood (EML) estimator or pseudo empirical maximum likelihood (PEML) estimator which has been the focus of most researches. The first reason concerns the interpretation of both estimators as minimizing a (directed) distance between the estimated probabilities p_i and the empirical frequencies $\frac{1}{n}$ or $\frac{d_i}{N}$. It seems appealing to weight the discrepancies using an efficient estimate of these probabilities (i.e., $\hat{p_i}$), as in KLEM procedure, rather than an inefficient estimate of these probabilities (i.e., $\frac{1}{n}$ or $\frac{d_i}{N}$), as in the EML or PEML procedure. The second reason concerns the relative robustness of the two estimators. The KLEM estimator is affected to a much lesser extent by perturbations in data (see Imbens, Spady and Johnson (1998)). The third, using empirical likelihood approach, the optimization problem fails to have a proper solution in some situations. However, the solution to our optimization problem exists and is unique. Furthermore, our method is a computationally simple approach.

In Section 2, we give a brief review of entropy and a justification for its use in finite populations, and explain how the usual KLEM approach relates to a linear model as the calibration estimator. In Section 3, we propose the model-calibration K-L relative entropy minimization (MKLEM) approach to incorporating auxiliary information into estimation of the population mean under a very general working model that includes linear and nonlinear regression and generalized linear models as special cases. We go on to show that the resulting estimator is asymptotically as efficient as model-calibration (MC) estimator. Our method is model-assisted, that is, the proposed estimators are asymptotically design unbiased irrespective of whether the working model is correct or not, and are particularly efficient if a working model is correct. In Section 4, we discuss estimation of the finite population distribution functions and show that our approach for mean case can be directly applied to estimation of the distribution function. We also show that our estimator be itself a distribution function and share the asymptotic efficiencies of a generalized regression estimator. Some proofs are given in Appendix.

2. The usual K-L relative entropy minimization approach

The importance of suitable measure 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. Since then various types of distance measures have been developed in literature. A concept closely related to the one of distance measures is that of divergence measure based on the idea of information-theoretic entropy first introduced in communication theory by Shannon (1949). Here we consider Shannon's concept of information-theoretic entropy and its generalization known as the Kullback and Leibler relative entropy or the divergence measure between two probability distributions. Any probability distribution $p_i, i = 1, 2, \cdots, n$ (say), of a random variable taking n values provides a measure of uncertainty regarding that random. In 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 we have in the probability distribution $p_i, i = 1, 2, \cdots, n$. 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 $-\log p_i, i = 1, 2, \cdots, n$, 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 (1), the K-L relative entropy of one probability distribution $\mathbf{p} = (p_1, p_2, \cdots, p_n)'$ with respect to another distribution $\mathbf{q} = (q_1, q_2, \cdots, q_n)'$ 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 the distance between two distributions. It is easy to see the link between $C(\mathbf{p}, \mathbf{q})$ and the Cressie and Read power divergence family (Bero and Bilias (2002)). If we choose $\mathbf{q} = (\frac{1}{n}, \frac{1}{n}, \cdots, \frac{1}{n})' = \frac{\mathbf{1}}{\mathbf{n}}$, where $\mathbf{1}$ is an $n \times 1$ vector of ones, $C(\mathbf{p}, \mathbf{q})$ reduces to

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

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

Consider a finite population, $\mathcal{S}$, consisting of N distinct units. Associated with the ith unit are the study variable y_i and a vector of auxiliary variables x_i. The values $x_1, x_2, \cdots, x_N$ are

known for the entire population (i.e., complete), but y_i is known only if the ith unit is selected in the samples. Assume the inclusion probabilities $\pi_i = pr$ $(i \in s)$ are strictly positive. For the moment we restrict attention to estimating the population mean $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i$. When no auxiliary information can be used, the conventional estimator of the finite population distribution function $F_N = N^{-1}\sum_{i=1}^{N}\delta_{z_i}$ is the Horvitz-Thompson estimator $\hat{F}_{HT} = \frac{1}{N}\sum_{i\in s} d_i\delta_{z_i}$, where δ_{z_i} is the point measure at z_i, and $d_i = 1/\pi_i$ are called the basic design weights. Now we know the finite population mean $\bar{X}$ of the auxiliary vector x, and the new estimator of F_N should be constructed as little as possible such that it modifies. A K-L relative entropy minimization estimator (KLEME) of $\bar{Y}_N$ is then defined by $\hat{\bar{Y}}_{KL} = \sum_{i\in s}\hat{p}_i y_i$, where $\hat{p}_i$'s are obtained by minimizing

$$C(\mathbf{p}, \mathbf{d}/\mathbf{N}) = \sum_{i\in s} p_i \log(Np_i/d_i) \tag{4}$$

subject to

$$\sum_{i\in s} p_i = 1, \qquad \sum_{i\in s} p_i x_i = \bar{X}, \qquad (0 \le p_i \le 1). \tag{5}$$

Using the Lagrange multiplier method, it is easily shown that

$$\hat{p}_i = \frac{d_i \exp\{\lambda u_i\}}{\sum_{j\in s} d_j \exp\{\lambda u_j\}}, \quad \text{for} \quad i \in s \tag{6}$$

with the lagrange multiplier, λ, is the solution to

$$\sum_{i\in s} d_i u_i \exp\{\lambda u_i\} = 0, \tag{7}$$

where $u_i = x_i - \bar{X}$ for $i = 1, 2, \cdots, N$. We have a similar result as Theorem 1 of Chen and Sitter (1999). For simplicity, we state results for a scalar x and y, though they hold generally.

Theorem 1 *Under conditions (i) and (ii) below, the KLEME of $\bar{Y}_N$, when $\bar{X}$ is known, is asymptotically equivalent to a generalized regression estimator (GREG). That is*

$$\hat{\bar{Y}}_{KL} = \sum_{i\in s} w_i[1 - \frac{(x_i - \bar{x}_w)(\bar{x}_w - \bar{X})}{\sum_{i\in s} w_i(x_i - \bar{x}_w)^2}]y_i + o_p(n^{-\frac{1}{2}}) = \bar{y}_{GREG} + o_p(n^{-\frac{1}{2}}), \tag{8}$$

where $\bar{x}_w = \sum_{i\in s} w_i x_i$, $w_i = d_i/\sum_{j\in s} d_j$, for $i \in s$.

The conditions needed are:

(i) $u^* = \max_{i\in s}|u_i| = o_p(n^{\frac{1}{2}})$ and (ii) $\sum_{i\in s} d_i u_i/\sum_{i\in s} d_i u_i^2 = O_p(n^{-\frac{1}{2}})$.

The point we want to illustrate is that, it is the relationship between y and x, hopefully captured by the working-model, that determines how the auxiliary information should best be used. In Theorem 1 we show that $\hat{\bar{Y}}_{KL}$, the KLEME of $\bar{Y}$, is asymptotically equivalent to the generalized regression estimator, $\bar{y}_{GREG}$, and the GREG is motivated as a model-assisted estimator using a linear working-model (Särndal, 1980). Thus, $\hat{\bar{Y}}_{KL}$ relies implicitly on a linear relationship between y and x without explicitly stating so. Another point relates to the issue of complete information on the x variable (i.e., known for all units in the population) versus only known the

value of its population mean, $\bar{X}$. The GREG is motivated by using the predicted values from a linear model for each $x_i, i = 1, \cdots, N$. However, the resulting estimator in (8) only needs $\bar{X}$ to be implemented. As we will see, this is related to the use of a linear model.

3. Estimation of mean under a general model

3.1. Model-calibration estimator

Let $u_i = u(x_i)$ be any known function of x_i. The usual calibration estimator (Deville and Särndal, 1992) of $\bar{Y}$ is constructed as $\hat{\bar{Y}}_C = \sum_{i\in s} w_i y_i$, where the calibration weights w_i 's are chosen to minimize their average distance Φ_s from the Horvitz-Thompson estimator $\hat{\bar{Y}}_{HT} = \frac{1}{N}\sum_{i\in s} d_i y_i$, subject to the constraint

$$\frac{1}{N}\sum_{i\in s} w_i = 1, \quad \sum_{i\in s} w_i u_i = \sum_{i=1}^{N} u_i. \tag{9}$$

The distance measure Φ_s is the chi-squared distance given by

$$\Phi_s = \sum_{i\in s}(w_i - d_i)^2/d_i. \tag{10}$$

Now suppose the relationship between y and x can be described by a super-population model through the first and second moments,

$$E_\xi(y_i|x_i) = \mu(x_i, \theta), \quad V_\xi(y_i|x_i) = v_i^2\sigma^2, \quad i = 1, 2, \cdots, N, \tag{11}$$

where $\theta = (\theta_0, \cdots, \theta_p)$ and σ^2 are unknown super-population parameters, $\mu(x, \theta)$ is a known function of x and θ, the v_i is a known function of x_i or $\mu_i = \mu(x_i, \theta)$, and E_ξ and V_ξ denote the expectation and variance with respect to the super-population model. We also assume that $(y_1, x_1), \cdots, (y_N, x_N)$ are mutually independent. The model structure (11) is quite general and includes two very important cases: (I) the linear or nonlinear regression model,

$$y_i = \mu(x_i, \theta) + v_i\varepsilon_i, \quad i = 1, 2, \cdots, N, \tag{12}$$

where ε_i's are independently and identically distributed random variables with $E_\xi(\varepsilon_i) = 0$ and $V_\xi(\varepsilon_i) = \sigma^2$, and $v_i = v(x_i)$ is a strictly positive known function of x_i only; and (II) the generalized linear model,

$$g(\mu_i) = x_i'\theta, \quad V_\xi(y_i|x_i) = v(\mu_i), i = 1, 2, \cdots, N, \tag{13}$$

where $\mu_i = E_\xi(y_i|x_i)$, and $g(\cdot)$ is a link function and $v(\cdot)$ is a variance function. The model-calibration approach can be accomplished by first using (y_i, x_i) for $i \in s$ to build the model and then calibrating over fitted values from the model. Consider a design-based method for estimating the model parameter θ. In this case, θ may be meaningless, and be replaced by θ_N, an estimate of θ based on the data from the entire population. θ_N is then estimated by $\hat{\theta}$, a design-based estimate from the sampled data through the general method of estimating

equations (Godmanbe and Thompson (1986), Wu and Sitter (2001)). The MC estimator $\hat{\bar{Y}}_{MC}$ is obtained by replacing u_i used in the constraint (9) with $\hat{\mu}_i = \mu(x_i, \hat{\theta})$. Using Lagrange multiplier approach, the resulting MC estimator is given by

$$\hat{\bar{Y}}_{MC} = \hat{\bar{Y}}_{HT} + \big\{N^{-1}\sum_{i=1}^{N}\hat{\mu}_i - N^{-1}\sum_{i\in s} d_i\hat{\mu}_i\big\}\hat{B}_N, \tag{14}$$

where $\hat{B}_N = \sum_{i\in s} d_i(\hat{\mu}_i - \bar{\mu})(y_i - \bar{y})/\sum_{i\in s} d_i(\hat{\mu}_i - \bar{\mu})^2$, $\bar{y} = \sum_{i\in s} d_i y_i/\sum_{i\in s} d_i$ and $\bar{\mu} = \sum_{i\in s} d_i\hat{\mu}_i/\sum_{i\in s} d_i$. If constraint $N^{-1}\sum_{i\in s} w_i = 1$ is dropped, the single calibration equation $\sum_{i\in s}\hat{\mu}_i = \sum_{i\in s} w_i\hat{\mu}_i$ yields

$$\hat{\bar{Y}}^{*}_{MC} = \hat{\bar{Y}}_{HT} + \big\{N^{-1}\sum_{i=1}^{N}\hat{\mu}_i - N^{-1}\sum_{i\in s} d_i\hat{\mu}_i\big\}\hat{B}^{*}_N, \tag{15}$$

where $\hat{B}^{*}_N = \sum_{i\in s} d_i\hat{\mu}_i y_i/\sum_{i\in s} d_i\hat{\mu}_i^2$.

3.2. Model-calibrated K-L relative entropy minimization estimator of the mean

To extend the K-L relative entropy minimization approach to model (11), we define scalar $u_i = u(y_i, x_i) = \mu(x_i, \hat{\theta}) - N^{-1}\sum_{i=1}^{N}\mu(x_i, \hat{\theta})$. The model-calibrated K-L relative entropy minimization estimator (MKLEME) of $\bar{Y}$ is then defined as $\hat{\bar{Y}}_{MKL} = \sum_{i\in s}\hat{p}_i y_i$, where $\hat{p}_i$'s minimize $C(\mathbf{p}, \mathbf{d}/\mathbf{N})$ subject to

$$\sum_{i\in s} p_i = 1, \quad \sum_{i\in s} p_i u_i = 0, \quad (0 \le p_i \le 1). \tag{16}$$

Using the Lagrange multiplier method we can easily get $\hat{p}_i = \frac{d_i \exp\{\lambda u_i\}}{\sum_{j\in s} d_j \exp\{\lambda u_j\}}$, for $i \in s$, and the scalar Lagrange multiplier λ is the solution to $\sum_{i\in s} d_i \exp\{\lambda u_i\} = 0$. Note that with no auxiliary information the resulting MKLEME of $\bar{Y}$ is $\hat{\bar{Y}} = \sum_{i\in s} d_i y_i/\sum_{i\in s} d_i$.

A theorem analogous to Theorem 2 of Wu and Sitter (2001) can then be proved. Assume that there is a sequence of sampling designs and a sequence of finite populations, indexed by ν. Both the sample size n_ν and the population size N_ν go to infinity as $\nu \to \infty$. All limiting processes are understood to be as $\nu \to \infty$, but the ν is suppressed to simplify notation. The conditions needed are:

(i) $\frac{1}{N}\sum_{i=1}^{N}|u_i|^3 = O(1)$, $\frac{1}{N}\sum_{i=1}^{N}|y_i|^3 = O(1)$;

(ii) $\sum_{i\in s} d_i u_i/\sum_{i\in s} d_i u_i^2 = O_p(n^{-\frac{1}{2}})$, $\frac{1}{N}\sum_{i\in s} d_i = 1 + o_p(n^{-\frac{1}{2}})$;

(iii) $\hat{\theta} = \theta_N + O_p(n^{-\frac{1}{2}})$, and $\theta_N \to \theta$;

(iv) For each $x_i, \partial\mu(x_i, t)/\partial t$ is continuous in t and $|\partial\mu(x_i, t)/\partial t| \le h(x_i, \theta)$ for t in a neighborhood of θ, $N^{-1}\sum_{i=1}^{N} h(x_i, \theta) = O(1)$ and $h^* = \max_{i\in s}|h_i| = o_p(n)$, where $h_i = h(x_i, \theta_N)$;

(v) The basic design weights, $d_i = 1/\pi_i$, satisfy that the Horvitz-Thompson estimator for certain population means is asymptotically normal distributed.

Theorem 2 *Under the same asymptotic framework and conditions (i)–(ii) above, we have*

$$\hat{\bar{Y}}_{MKL} = \hat{\bar{Y}}_{MC} + o_p(n^{-\frac{1}{2}}). \tag{17}$$

From Theorem 2 above and Theorem 1 in Wu and Sitter (2001), we can summarize one important property of $\hat{\bar{Y}}_{MKL}$ in the following corollary.

Corollary *Under the same conditions (i)–(v), we have*

$$\hat{\bar{Y}}_{MKL} = \hat{\bar{Y}}_{HT} + O_p(n^{-\frac{1}{2}}), \tag{18}$$

which is thus asymptotically design-unbiased estimators for $\bar{Y}$, irrespective of whether the model is correct or not, and is approximately model-unbiased.

4. Estimation of the finite population distribution function

The finite population distribution function $F_Y(y) = N^{-1}\sum_{i=1}^{N} I(y_i \le y)$ is also a finite population mean of an indicator variable $z_i = I(y_i \le y)$. Without using any auxiliary information, estimation of the distribution function is a special case of the population mean and is usually straightforward. In the presence of the auxiliary information, there exist several general estimation procedures in recent literature to obtain more efficient estimators for the population mean and total, which have been introduced in Section 1. If we try to directly apply these general procedures to the estimation of the distribution function, however, due to the specific nature of a distribution function, the resulting estimators often have some undesirable properties. The estimation of the distribution function using auxiliary information differs from the estimation of the population mean in several fundamental aspects (Chen and Wu, 2002). The distribution function involves a dichotomous variable $I(y_i \le y)$. We need special treatment for the modelling process to obtain efficient estimators for the distribution function. Also, it is desirable to require that estimators of the distribution function are themselves distribution functions. Quantile estimates can then be obtained by direct inversion of the estimated distribution function. An attractive advantage of our KLEM approach is the intrinsic properties of the resulting weights: $\hat{p}_i > 0$ and $\sum_{i\in s}\hat{p}_i = 1$, which make this approach generally applicable to estimation of distribution functions and quantiles.

To estimate $F_Y(y)$ for a given y_0 we need to replace y_i by $z_i = I(y_i \le y_0)$. The optimal choice of u_i in (16) is given by $u_i = E_\xi(z_i|x_i) = P(y_i \le y_0|x_i)$. It is important to notice that the optimal choice of u_i depends on y_0. No u_i with a fixed y_0 can be uniformly optimal for all values of y.

A commonly used working model for the finite population is the regression model given in (12): $y_i = \mu(x_i,\theta) + v_i\varepsilon_i$, $i = 1, 2, \cdots, N$. Let θ_N and σ_N be the estimators of θ and σ based on data from the entire finite population, respectively. Under model (12), $E_\xi(z_i|x_i) = P(y_i \le y|x_i) = G[\{y - \mu(x_i,\theta)\}/v_i]$, where $G(\cdot)$ is the cumulative distribution function (cdf) of the error term. In many situations, it is reasonable to assume that the error term ε_i in model (12) are normally distributed. In this case , let $u_i = u_i(\theta_N, \sigma_N, y) = \Phi[\{y - \mu(x_i,\theta_N)\}/v_i\sigma_N]$, where $\Phi(\cdot)$

is the cdf of standard normal distribution. Note that θ_N, not θ, is used for defining u_i. With this treatment, u_i is well defined over the finite population and this makes all design-based arguments possible. Let $\hat{\theta}$ and $\hat{\sigma}$ be design-based estimates for θ_N and σ_N, respectively.

The MKLEM estimator of $F_Y(y)$ is defined as

$$\hat{F}_{MKL}(y) = \sum_{i\in s} \hat{p}_i z_i = \sum_{i\in s} \hat{p}_i I(y_i \le y),$$

where the $\hat{p}_i$'s minimize $C(\mathbf{p}, \mathbf{d}/\mathbf{N})$ subject to

$$\sum_{i\in s} p_i = 1, \quad \sum_{i\in s} p_i u_i(\hat{\theta}, \hat{\sigma}, y_0) = N^{-1} \sum_{i=1}^{N} u_i(\hat{\theta}, \hat{\sigma}, y_0) \quad (0 \le p_i \le 1). \tag{19}$$

The value of y_0 is pre-specified.

Let $\bar{z}_d = (\sum_{i\in s} d_i)^{-1} \sum_{i\in s} d_i z_i$, $\bar{u}_d = (\sum_{i\in s} d_i)^{-1} \sum_{i\in s} d_i u_i(\theta_N, \sigma_N, y_0)$ and $\bar{U}_N = N^{-1}$ $\sum_{i=1}^{N} u_i(\theta_N, \sigma_N, y_0)$. The following result is then established.

Theorem 3 *Under the conditions below, we have*

(1) the model-calibrated K-L relative entropy minimization estimator $\hat{F}_{MKL}(y)$ is asymptotically equivalent to a generalized regression estimator, i.e.

$$\hat{F}_{MKL}(y) = \bar{z}_d + (\bar{U}_N - \bar{u}_d) B_N + o_p(n^{-\frac{1}{2}}),$$

where $B_N = \sum_{i=1}^{N} \{u_i(\theta_N, \sigma_N, y_0) - \bar{U}_N\} z_i / \sum_{i=1}^{N} \{u_i(\theta_N, \sigma_N, y_0) - \bar{U}_N\}^2$.

(2) $\hat{F}_{MKL}(y)$ is asymptotically design-consistent estimator of $F_Y(y)$ and is also approximately model-unbiased under (12) at $y = y_0$.

The conditions needed are:

(a) $|\hat{\theta} - \theta_N| = O_p(n^{-\frac{1}{2}})$ and $\hat{\sigma} - \sigma_N = O_p(n^{-\frac{1}{2}})$;

(b) $\max_{i\in s} n d_i / N = O(1)$;

(c) For each x_i, $\mu(x_i, \theta)$ is twice differentiable,

$$N^{-1} \sum_{i=1}^{N} \{\mu'(x_i, \theta_N)/v_i\}^2 = O(1), \quad N^{-1} \sum_{i=1}^{N} \{\mu(x_i, \theta_N)/v_i\}^2 = O(1)$$

and $N^{-1} \sum_{i=1}^{N} v_i^{-2} = O(1)$.

Note that y_0 used in (19) is fixed, the weights $\hat{p}_i'$s are independent of y. It is easy to see that $\hat{F}_{MKL}(y)$ is itself a distribution function. $\hat{F}_{MKL}(y)$ will be most efficient at y in the neighborhood of y_0. The value of y_0 can be easily specified according to efficiency considerations.

Appendix: Some proofs

Proof of Theorem 1 It is easy to show that the solution λ of Equation (7) exists. Then we have

$$\begin{aligned} 0 &= \sum_{i\in s} d_i u_i \exp\{\lambda u_i\} = \sum_{i\in s} d_i u_i [1 + \lambda u_i + o(\lambda u_i)] \\ &= \sum_{i\in s} d_i u_i + \lambda \cdot \sum_{i\in s} d_i u_i^2 + \lambda \cdot o(\sum_{i\in s} d_i u_i^2). \end{aligned} \tag{A.1}$$

Therefore, we have

$$\lambda = -\frac{\sum_{i\in s} w_i u_i}{\sum_{i\in s} w_i u_i^2} + o_p(n^{-\frac{1}{2}}) = O_p(n^{-\frac{1}{2}}). \tag{A.2}$$

Using the Taylor expansion, (A.2) and conditions (i) and (ii), we have

$$\sum_{i\in s} d_i \exp\{\lambda u_i\} = \sum_{i\in s}[d_i(1+\lambda u_i + o(\lambda u_i))] = (\sum_{i\in s} d_i)[1+\lambda \bar{u}_w + o_p(n^{-\frac{1}{2}})]. \tag{A.3}$$

Note that here $u_i = x_i - \bar{X}$, it is easy to find that

$$\begin{aligned}
\hat{\bar{Y}}_{KL} &= \sum_{i\in s} \hat{p}_i y_i = (\sum_{i\in s} d_i \exp\{\lambda u_i\})^{-1}(\sum_{i\in s} d_i y_i \exp\{\lambda u_i\}) \\
&= [\sum_{i\in s} w_i y_i + \lambda \sum_{i\in s} w_i y_i u_i + o_p(n^{-\frac{1}{2}})][1+\lambda \bar{u}_w + o(\lambda \bar{u}_w)]^{-1} \\
&= [\sum_{i\in s} w_i y_i + \lambda \sum_{i\in s} w_i y_i u_i + o_p(n^{-\frac{1}{2}})][1-\lambda \bar{u}_w + o_p(n^{-\frac{1}{2}})] \\
&= \sum_{i\in s} w_i [1 - \frac{(x_i - \bar{x}_w)(\bar{x}_w - \bar{X})}{\sum_{i\in s} w_i (x_i - \bar{x}_w)^2}] y_i + o_p(n^{-\frac{1}{2}}) \\
&= \bar{y}_{GREG} + o_p(n^{-\frac{1}{2}}).
\end{aligned}$$

Proof of Theorem 2 Similar to the proof of Theorem 1, we must have

$$\lambda = -\frac{\sum_{i\in s} d_i u_i}{\sum_{i\in s} d_i u_i^2} + o_p(n^{-\frac{1}{2}}) = O_p(n^{-\frac{1}{2}}). \tag{A.4}$$

Using the Taylor expansion, (A.4) and condition (i), we have

$$\sum_{i\in s} d_i \exp\{\lambda u_i\} = \sum_{i\in s}[d_i(1+\lambda u_i + o(\lambda u_i))] = (\sum_{i\in s} d_i)[1+\lambda \frac{\sum_{i\in s} d_i u_i}{\sum_{i\in s} d_i} + o_p(n^{-\frac{1}{2}})]. \tag{A.5}$$

From (A.4) and the conditions in Theorem 2 we get

$$\begin{aligned}
\hat{\bar{Y}}_{MKL} &= \sum_{i\in s} \hat{p}_i y_i = (\sum_{i\in s} d_i \exp\{\lambda u_i\})^{-1}(\sum_{i\in s} d_i y_i \exp\{\lambda u_i\}) \\
&= [\frac{1}{N}\sum_{i\in s} d_i y_i + \lambda \frac{1}{N}\sum_{i\in s} d_i y_i u_i + o_p(n^{-\frac{1}{2}})][1+\lambda \bar{u}_w + o(\lambda \bar{u}_w)]^{-1} \\
&= [\frac{1}{N}\sum_{i\in s} d_i y_i + \lambda \frac{1}{N}\sum_{i\in s} d_i y_i u_i + o_p(n^{-\frac{1}{2}})][1-\lambda \frac{\sum_{i\in s} d_i u_i}{\sum_{i\in s} d_i} + o_p(n^{-\frac{1}{2}})] \\
&= \hat{\bar{Y}}_{HT} + \{\frac{1}{N}\sum_{i=1}^{N} u_i - \frac{1}{N}\sum_{i\in s} d_i u_i\} \frac{\sum_{i\in s} d_i (y_i - \bar{y}_w)(u_i - \bar{u}_w)}{\sum_{i\in s} d_i (u_i - \bar{u}_w)^2} + o_p(n^{-\frac{1}{2}}) \\
&= \hat{\bar{Y}}_{MC} + + o_p(n^{-\frac{1}{2}}).
\end{aligned} \tag{A.6}$$

Then, we obtain the result of Theorem 2.

Proof of Theorem 3 Proof of part (1): Suppose that θ_N and σ_N are known, and use $u_i =$

$u_i(\theta_N, \sigma_N, y_0)$ in the constraint (19) to construct $\hat{F}_{MKL}(y)$. Let $g_i = u_i(\theta_N, \sigma_N, y_0) - \bar{U}_N$. It follows from the proof of Theorem 1 that

$$\hat{p}_i = \frac{d_i \exp\{\lambda g_i\}}{\sum_{j\in s} d_j \exp\{\lambda g_j\}} = \frac{w_i}{1+\lambda g_i} + o_p(n^{-\frac{1}{2}}) = w_i(1-\lambda g_i) + o_p(n^{-\frac{1}{2}})$$

and

$$\lambda = -s_{wg}^{-2}\bar{g}_w + + o_p(n^{-\frac{1}{2}}),$$

with $w_i = d_i/\sum_{i\in s} d_i$, $\bar{g}_w = \sum_{i\in s} w_i g_i$, $s_{wg}^2 = \sum_{i\in s} w_i g_i^2$. Hence, it follows from the above expansion that

$$\hat{F}_{MKL}(y) = \bar{z}_d + (\bar{U}_N - \bar{u}_d)B_N + o_p(n^{-\frac{1}{2}}).$$

When $u_i = u_i(\theta_N, \sigma_N, y_0)$ is replaced by $\hat{u}_i = u_i(\hat{\theta}, \hat{\sigma}, y_0)$, we need only to show that

$$\frac{1}{N}\sum_{i=1}^{N}\hat{u}_i - (\sum_{i\in s} d_i)^{-1}\sum_{i\in s} d_i\hat{u}_i = \frac{1}{N}\sum_{i=1}^{N} u_i - (\sum_{i\in s} d_i)^{-1}\sum_{i\in s} d_i u_i + o_p(n^{-\frac{1}{2}}). \tag{A.7}$$

Condition (b) implies that $N^{-1}\sum_{i\in s} d_i c_i - N^{-1}\sum_{i=1}^{N} c_i = O_p(n^{-\frac{1}{2}})$ if $N^{-1}\sum_{i=1}^{N} c_i^2 = O(1)$. Let $a_i(\theta,\sigma) = (\partial/\partial\theta)u_i(\theta,\sigma), b_i(\theta,\sigma) = (\partial/\partial\sigma)u_i(\theta,\sigma)$. It follows from a Taylor series expansion that

$$\hat{u}_i = u_i + [a_i(\theta_N,\sigma_N)]'(\hat{\theta}-\theta_N) + b_i(\theta_N,\sigma_N)(\hat{\sigma}-\sigma_N) + O_p(\|\hat{\theta}-\theta_N\|) + O_p((\hat{\sigma}-\sigma_N)^2).$$

Under the regularity conditions, we have

$$\frac{1}{N}\sum_{i=1}^{N} a_i - (\sum_{i\in s} d_i)^{-1}\sum_{i\in s} d_i a_i = O_p(n^{-\frac{1}{2}}), \quad \frac{1}{N}\sum_{i=1}^{N} b_i - (\sum_{i\in s} d_i)^{-1}\sum_{i\in s} d_i b_i = O_p(n^{-\frac{1}{2}}).$$

This together with (a) proves the first part of Theorem 3.

Proof of part (2): From part (1), $\hat{F}_{MKL}(y) = F_{HT}(y) + O_p(n^{-\frac{1}{2}})$, so $\hat{F}_{MKL}(y)$ is asymptotically design-consistent. Assume $(x_i, y_i), i = 1, \cdots, N$, are iid random variates from the superpopulation. Then we have $\theta_N = \theta + O_p(n^{-\frac{1}{2}})$. To prove that $\hat{F}_{MKL}(y)$ is approximately model-unbiased at $y = y_0$, we substitute $\hat{u}_i = u_i(\hat{\theta}, \hat{\sigma}, y_0)$ used in constraint (19) by $u_i^* = u_i(\theta, \sigma, y_0)$, and denote the resulting estimator by $\hat{F}^*_{MKL}(y) = \sum_{i\in s} p_i^* I(y_i \le y)$. Under model (12), we have

$$E_\xi\{\hat{F}^*_{MKL}(y_0)\} = \sum_{i\in s} p_i^* E_\xi\{I(y_i \le y_0)\} = \sum_{i\in s} p_i^* u_i^* = \frac{1}{N}\sum_{i=1}^{N} u_i^* = \frac{1}{N}\sum_{i=1}^{N} u_i^* = E_\xi\{F_Y(y_0)\}.$$

Hence $\hat{F}^*_{MKL}(y_0)$ is exactly model-unbiased for $F_Y(y_0)$. However, there is a conceptual gap between $\hat{\theta}$ and θ: It is usually true that $\hat{\theta} = \theta_N + O_p(n^{-\frac{1}{2}})$ under the design-based framework and $\theta_N = \theta + O_p(n^{-\frac{1}{2}})$ under the superpopulation model. To conclude that $\hat{\theta} = \theta + O_p(n^{-\frac{1}{2}})$ we need to consider the joint expectation under both the design and the model. Suppose that $\hat{\theta} = \theta + O_p(n^{-\frac{1}{2}})$ under the model. It is then straightforward to show that $\hat{u}_i = u_i^* + o_p(1)$, with the uniform term $o_p(1)$ over i. Similarly, $g_i = g_i^* + o_p(1)$ uniformly over i, where $g_i =$

$\hat{u}_i - N^{-1}\sum_{i=1}^{N}\hat{u}_i$ and $g_i^* = u_i^* - N^{-1}\sum_{i=1}^{N}u_i^*$. It now follows that $\hat{p}_i = w_i/(1+\lambda g_i) + o_p(n^{-\frac{1}{2}}) = w_i/(1+\lambda g_i^*) + o_p(1) = p_i^* + o_p(1)$. The approximate model unbiasedness follows from $\hat{F}_{MKL}(y_0) = \hat{F}^*_{MKL}(y_0) + o_p(1)$.

References:

[1] BERO A K, BILIAS Y. *The MM, ME, ML, EL, EF and GMM approaches to estimation: a synthesis* [J]. J. Econometrics, 2002, **107**: 51–86.

[2] CASSEL C E, SÄRNDAL C E, WRETMAN J H. *Some results on generalized difference estimation and generalized regression estimation for finite populations* [J]. Biometrika, 1976, **63**: 615–620.

[3] CHEN Jia-hua, QIN Jin. *Empirical likelihood estimation for finite populations and the effective usage of auxiliary information* [J]. Biometrika, 1993, **80**: 107–116.

[4] CHEN Jia-hua, SITTER R R. *A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys* [J]. Statist. Sinica, 1999, **9**: 385–406.

[5] CHEN Jia-hua, WU Chang-bao. *Estimation of distribution function and quantiles using the model-calibrated pseudo empirical likelihood method* [J]. Statist. Sinica, 2002, **12**: 1223–1239.

[6] DEVILLE J C, SÄRNDAL C E. *Calibration estimators in survey sampling* [J]. J. Amer. Statist. Assoc., 1992, **87**: 376–382.

[7] GODMANBE V P, THOMPSON M E. *Parameters of superpopulation and survey population: Their relationshipa and estimation* [J]. Int. Statist. Rev., 1986, **54**: 127–138.

[8] IMBENS G W, SPADY R H, JOHNSON P. *Information theoretic approaches to inference in moment condition models* [J]. Econometrica, 1998, **66**: 333–357.

[9] SÄRNDAL C E. *On π-inverse weighting versus best linear unbiased weighting in probability sampling* [J]. Biometrika, 1980, **67**: 639–650.

[10] SHANNON C E, WEAVER W. *The Mathematical Theory of Communication* [M]. University of Illinois Press, Urbana, 1949.

[11] ZHONG B, RAO J N K. *Empirical likelihood inference under stratified random sampling using auxiliary population information* [J]. Biometrika, 2000, **87**: 929–938.

[12] WU Chang-bao, SITTER R R. *A model-calibration approach to using complete auxiliary information from survey data* [J]. J. Amer. Statist. Assoc., 2001, **96**: 185–193.

利用完全辅助信息的模型校正信息论方法

伍长春[1,2], 张润楚[2]

(1. 嘉兴学院数学与信息科学学院, 浙江 嘉兴 314001;
2. 南开大学数学科学学院统计系和教育部重点实验室, 天津 300071)

摘要: 本文我们提出了使用调查数据中完全辅助信息的模型校正 K-L 相对熵最小化方法. 在估计有限总体均值时我们的估计渐近等价于 MC 估计 (Wu and Sitter (2001)). 我们方法一个有吸引力的优点是, 导出的权具有特征: $\hat{p}_i > 0$ 和 $\sum_{i\in s}\hat{p}_i = 1$. 这使得可把此方法应用于估计分布函数和分位数. 导出的分布函数估计量 $\hat{F}_{MKL}(y)$ 渐近等价于广义回归估计, 且本身是一分函数布.

关键词: 模型校正; 完全辅助信息; K-L 相对熵; 广义回归估计; 经验似然