Statistica Sinica **12**(2002), 1223-1239

ESTIMATION OF DISTRIBUTION FUNCTION AND QUANTILES USING THE MODEL-CALIBRATED PSEUDO EMPIRICAL LIKELIHOOD METHOD

Jiahua Chen and Changbao Wu

University of Waterloo

Abstract: We use the model-calibrated pseudo empirical likelihood method to construct estimators for the finite population distribution function. Under an assumed superpopulation working model, the proposed estimators have minimum model expectation of asymptotic design-based variance among a class of estimators and therefore are optimal in that class. The estimators are asymptotically design-unbiased irrespective of the working model and are also approximately model-unbiased under the model. They share the design-based asymptotic efficiency with that of a generalized regression estimator but, unlike the latter, the estimators are genuine distribution functions. Quantile estimation through direct inversion and using a model-calibrated difference estimator are studied, and their asymptotic efficiency is investigated through Bahadur representations. Variance estimation and confidence intervals for the distribution function are also addressed. Results of a limited simulation study regarding the finite sample performance of proposed estimators are reported.

Key words and phrases: Auxiliary information, Bahadur representation, design consistency, finite population, model-assisted approach, model calibration, variance estimation.

1. Introduction

The finite population distribution function $F_Y(y) = N^{-1}\sum_{i=1}^{N} I(y_i \leq y)$ is also a finite population mean of an indicator variable $z_i = I(y_i \leq y)$. Here y_i is the value of the study variable Y attached to the ith unit, $z_i = 1$ if $y_i \leq y$ and $z_i = 0$ otherwise. 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 auxiliary information, there exist several general estimation procedures in recent literature to obtain more efficient estimators for the population means and totals. Effort has been made to directly apply these general procedures for 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 model-based prediction estimator $\hat{F}_m(y)$ proposed by Chambers and Dunstan (1986) stimulated much of the later work in this area. Their proposed estimator $\hat{F}_m(y)$ is model-unbiased, i.e., $E_\xi\{\hat{F}_m(y) - F_Y(y)\} = 0$, where E_ξ denotes the expectation under the assumed superpopulation model ξ. But $\hat{F}_m(y)$ is design-inconsistent, i.e., $E_p\{\hat{F}_m(y)\} - F_Y(y)$ does not converge to zero as $N \to \infty$, where E_p denotes the expectation under the sampling design p. Careful model checking and diagnostics need to be carried out before this purely model-based estimator is used.

Rao, Kovar and Mantel (1990) proposed a design-based difference estimator $\hat{F}_d(y)$. However, $\hat{F}_d(y)$ may take values out of the range [0, 1] and it is not always a monotone function of y. We cannot always invert $\hat{F}_d(y)$ to obtain estimates for the population quantiles. Rao *et al.* (1990) suggest transforming $\hat{F}_d(y)$ into a monotone function before obtaining estimates for quantiles. The same idea was also discussed by Francisco and Fuller (1991). The implementation of such a process is not trivial and the loss of efficiency during the process is unknown.

Under simple random sampling, Wang and Dorfman (1996) proposed a hybrid estimator $\hat{F}_w(y)$ which is a weighted average of $\hat{F}_m(y)$ and $\hat{F}_d(y)$. Under certain conditions the new estimator is more efficient than both $\hat{F}_d(y)$ and $\hat{F}_m(y)$. However, it inherits the drawbacks of both estimators and the estimator cannot be readily generalized to more complex sampling designs.

The generalized regression estimator (GREG) for the population means or totals is the most popular one to use among survey practitioners. When this technique is directly applied to the distribution function, the resulting estimator shares many of the drawbacks of $\hat{F}_d(y)$. To estimate the distribution function having the same efficiency as GREG but without these limitations becomes an interesting problem.

Estimation of the distribution function using auxiliary information differs from the estimation of the population mean in several fundamental aspects. Most existing approaches for the estimation of population means or totals are either model-based or model-assisted with models directly specified over the study variables. The distribution function involves a dichotomous variable $I(y_i \le y)$. We need special treatment for the modeling process to obtain efficient estimators for the distribution function. Also, it is desirable to require that estimators of the distribution function be themselves distribution functions. Quantile estimates can then be obtained by direct inversion of the estimated distribution function.

In this article, we propose estimators for the finite population distribution function and its quantiles using the recently developed model-calibration and pseudo empirical likelihood methods (Chen and Sitter (1999); Wu and Sitter (2001a)). The proposed estimators effectively use auxiliary information at the estimation stage and possess a number of attractive features. Under the assumed working model, our proposed estimators are optimal in the sense of minimum model expectation of the design-based asymptotic variance. They are asymptotically design-unbiased irrespective of the working model and are also approximately model-unbiased under the model. In Section 2, we construct estimators for the distribution function under three different scenarios using the model-calibrated pseudo empirical likelihood method. The proposed estimators share the design-based asymptotic efficiency with that of a generalized regression estimator and are genuine distribution functions. Quantile estimation through direct inversion of the estimated distribution function is discussed in Section 3 and its asymptotic efficiency is studied through Bahadur representations. Also in Section 3, we propose a model-calibrated difference estimator for the quantiles that works well for both a general sampling design and a general working model. Variance estimation and confidence intervals for the distribution function are addressed in Section 4. Results of a limited simulation study regarding the finite sample performance of proposed estimators are reported in Section 5. Some concluding remarks are given in Section 6. All proofs are deferred to the Appendices.

2. Estimation of the Finite Population Distribution Function

The logical connection between the estimation of the population mean and the estimation of the distribution function using auxiliary information can be best seen under the model calibration and the model-calibrated pseudo empirical likelihood approaches (Wu and Sitter (2001a)). Let $(y_i, \boldsymbol{x}_i)$ be the values of the study variable Y and the vector auxiliary variable $\boldsymbol{X}$ for the ith unit in the finite population, $i = 1, \ldots, N$. We assume the values of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ are known. Let s be the set of units included in the sample.

Let $g_i = g(\boldsymbol{x}_i)$ be any known function of $\boldsymbol{x}_i$. The ordinary pseudo empirical maximum likelihood estimator (Chen and Sitter (1999)) of the finite population mean $\bar{Y}$ is defined as $\hat{\bar{Y}} = \sum_{i\in s} \hat{p}_i y_i$, where the $\hat{p}_i$'s maximize the pseudo empirical log-likelihood function

$$\hat{l}(\boldsymbol{p}) = \sum_{i\in s} d_i \log p_i, \tag{1}$$

subject to

$$\sum_{i\in s} p_i = 1, \quad \sum_{i\in s} p_i g_i = \frac{1}{N}\sum_{i=1}^{N} g_i \quad (0 < p_i < 1), \tag{2}$$

where $d_i = 1/\pi_i$ and $\pi_i = P(i \in s)$ are the first order inclusion probabilities.

The model-calibrated pseudo empirical maximum likelihood (MPEML) estimator $\hat{\bar{Y}}_{EL}$ (Wu and Sitter (2001a)) is obtained by first adapting a superpopulation model and then calibrating over-fitted values from the model. Suppose the relationship between Y and $\boldsymbol{X}$ can be depicted through a semi-parametric model: $E_\xi(y_i|\boldsymbol{x}_i) = \mu(\boldsymbol{x}_i, \boldsymbol{\theta})$, $V_\xi(y_i|\boldsymbol{x}_i) = \sigma_i^2$, $i = 1, \ldots, N$, where E_ξ, V_ξ denote the expectation and variance with respect to the superpopulation model, and $\boldsymbol{\theta}$ is a vector of parameters of the superpopulation. A design-based estimator $\hat{\boldsymbol{\theta}}$ for the model parameter $\boldsymbol{\theta}$ can be obtained through the general method of estimating equations (Godambe and Thompson (1986); Wu and Sitter (2001a)). The MPEML estimator $\hat{\bar{Y}}_{EL}$ is obtained by replacing g_i used in the constraint (2) with $g_i(\boldsymbol{\theta}) = E_\xi(y_i|\boldsymbol{x}_i) = \mu(\boldsymbol{x}_i, \boldsymbol{\theta})$. This choice of g_i is optimal in that the resulting estimator $\hat{\bar{Y}}_{EL}$ has minimum model expectation of design-based asymptotic variance among a class of estimators (Wu (2001)). Despite the underlying non-parametric likelihood motivations, $\hat{\bar{Y}}_{EL}$ is asymptotically equivalent to the generalized regression estimator under a linear regression model (Wu and Sitter (2001a)). The most attractive feature of $\hat{\bar{Y}}_{EL}$, however, is the intrinsic properties of the weights: $\hat{p}_i > 0$ and $\sum_{i\in s} \hat{p}_i = 1$. This will play a key role in the following development.

To estimate $F_Y(y)$ for a given y_0 we need to replace y_i by $z_i = I(y_i \le y_0)$. Among all choices of $g_i = g(\boldsymbol{x}_i)$ in (2), $g_i = E_\xi(z_i|\boldsymbol{x}_i) = P(y_i \le y_0|\boldsymbol{x}_i)$ minimizes the model expectation of design-based asymptotic variance of the resulting estimator for $F_Y(y)$. It is important to notice that this optimal choice of g_i depends on y_0. No g_i with a fixed y_0 can be uniformly optimal for $F_Y(y)$ for all values of y. Also, since the "response variable" is $z_i = I(y_i \le y)$, two types of working models can be considered: models that relate the y_i to the $\boldsymbol{x}_i$ or models that relate the indicator variable $I(y_i \le y)$ to the $\boldsymbol{x}_i$. The g_i's are called the fitted values for the z_i's. In what follows, we propose three MPEML estimators based on three different working models.

2.1. Obtain fitted values from a regression model

A commonly used working model for the finite population is

$$y_i = \mu(\boldsymbol{x}_i, \boldsymbol{\theta}) + v_i\varepsilon_i\,, \quad i = 1, \ldots, N, \tag{3}$$

where v_i is a known function of $\boldsymbol{x}_i$, and ε_i's are independent and identically distributed (i.i.d.) random variables with mean 0 and variance σ^2. Hence, we can model z_i indirectly through a model for y_i. For a linear regression model $\mu(\boldsymbol{x}_i, \boldsymbol{\theta}) = \boldsymbol{x}_i'\boldsymbol{\theta}$, but other non-linear models can also be considered. Let $\boldsymbol{\theta}_N$ and σ_N be respectively the estimators of $\boldsymbol{\theta}$ and σ based on data from the entire finite population. For instance, under a linear regression model with homogeneous variance and $\boldsymbol{\theta}$ of dimension p, we have $\boldsymbol{\theta}_N = (\boldsymbol{X}_N'\boldsymbol{X}_N)^{-1}\boldsymbol{X}_N'\boldsymbol{Y}_N$, where $\boldsymbol{X}_N$ is the $N \times p$ matrix with rows $\boldsymbol{x}_i'$ for $i = 1, \ldots, N$, $\boldsymbol{Y}_N = (y_1, \ldots, y_N)'$, and $\sigma_N^2 = (\boldsymbol{Y}_N - \boldsymbol{X}_N\boldsymbol{\theta}_N)'(\boldsymbol{Y}_N - \boldsymbol{X}_N\boldsymbol{\theta}_N)/(N-p)$.

Under (3), $E_\xi(z_i|\boldsymbol{x}_i) = P(y_i \leq y|\boldsymbol{x}_i) = G[\{y - \mu(\boldsymbol{x}_i, \boldsymbol{\theta})\}/v_i]$, where $G(\cdot)$ is the cumulative distribution function (cdf) of the error term, ε_i. In many situations it is reasonable to assume that the error terms ε_i in model (3) are normally distributed. In this case $g_i = g_i(\boldsymbol{\theta}_N, \sigma_q N, y) = \Phi[\{y - \mu(\boldsymbol{x}_i, \boldsymbol{\theta}_N)\}/(v_i\sigma_N)]$, where $\Phi(\cdot)$ is the cdf of standard normal distribution. Note that $\boldsymbol{\theta}_N$, not $\boldsymbol{\theta}$, is used in defining g_i. With this treatment g_i is well defined over the finite population and this makes all design-based arguments possible. Let $\hat{\boldsymbol{\theta}}$ and $\hat{\sigma}$ be design-based estimates for $\boldsymbol{\theta}_N$ and σ_N, respectively.

We define our first MPEML estimator of $F_Y(y)$ as $\hat{F}_{EL}^{(1)}(y) = \sum_{i\in s} \hat{p}_i z_i = \sum_{i\in s} \hat{p}_i I(y_i \leq y)$, where the $\hat{p}_i$'s maximize $\hat{l}(\boldsymbol{p})$ subject to

$$\sum_{i\in s} p_i = 1, \quad \sum_{i\in s} p_i g_i(\hat{\boldsymbol{\theta}}, \hat{\sigma}, y_0) = \frac{1}{N}\sum_{i=1}^{N} g_i(\hat{\boldsymbol{\theta}}, \hat{\sigma}, y_0) \quad (0 < p_i < 1)\,. \tag{4}$$

The value of y_0 is pre-specified. Let $\bar{z}_d = (\sum_{i\in s} d_i)^{-1} \sum_{i\in s} d_i z_i$, $\bar{g}_N = N^{-1} \sum_{i=1}^{N} g_i(\boldsymbol{\theta}_N, \sigma_N, y_0)$, and $\bar{g}_d = (\sum_{i\in s} d_i)^{-1} \sum_{i\in s} d_i g_i(\boldsymbol{\theta}_N, \sigma_N, y_0)$. Note that $\bar{z}_d$ is the usual Horvitz-Thompson estimator for $F_Y(y)$. The following results have been established. The proof is given in Appendix 1.

Theorem 1. (1) *Under the regularity conditions specified in Appendix* 1, *the model-calibrated pseudo empirical maximum likelihood estimator* $\hat{F}_{EL}^{(1)}(y)$ *is asymptotically equivalent to a generalized regression type estimator:*

$$\hat{F}_{EL}^{(1)}(y) = \bar{z}_d + (\bar{g}_N - \bar{g}_d)B_N + o_p(n^{-1/2})\,, \tag{5}$$

where $B_N = \sum_{i=1}^{N}\{g_i(\boldsymbol{\theta}_N, \sigma_N, y_0) - \bar{g}_N\}z_i / \sum_{i=1}^{N}\{g_i(\boldsymbol{\theta}_N, \sigma_N, y_0) - \bar{g}_N\}^2$.
(2) $\hat{F}_{EL}^{(1)}(y)$ *is asymptotically design-consistent estimator of* $F_Y(y)$ *and is also approximately model-unbiased under* (3) *at* $y = y_0$.

Note that the generalized regression type estimator referred to in (5) is for asymptotic comparison only. It is not a real estimator. Also note that y_0 used in (4) is fixed, the weights $\hat{p}_i$'s are independent of y. It is easy to see that $\hat{F}_{EL}^{(1)}(y)$ is itself a distribution function. Hence, $\hat{F}_{EL}^{(1)}(y)$ shares the asymptotic efficiencies of a generalized regression type estimator without the drawbacks of those of $\hat{F}_d(y)$. However, $\hat{F}_{EL}^{(1)}(y)$ 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.

The normality assumption about the error terms ε_i is used in proving (13) of Appendix 1 where a Taylor series expansion has been applied to $\hat{g}_i = \Phi[\{y - \mu(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}})\}/(v_i\hat{\sigma})]$. When this assumption is not desirable, the cdf of ε_i, $G(\cdot)$, will have to be estimated from the fitted residuals. Let $\hat{\varepsilon}_i = \{y_i - \mu(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}})\}/v_i$ for $i = 1, \ldots, N$, $G_n(u) = \sum_{i\in s} d_i I(\hat{\varepsilon}_i \leq u) / \sum_{i\in s} d_i$. The g_i used in constraint (4) can then be replaced by $g_i = G_n[\{y_0 - \mu(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}})\}/v_i]$. Under some mild regularity conditions, Theorem 1 and the variance estimator for $\hat{F}_{EL}(y)$ presented in Section 4.1 can still be used when $G(\cdot)$ is replaced by $G_n(\cdot)$. A key argument in proving this is to show that replacing $\hat{\boldsymbol{\theta}}$ by $\boldsymbol{\theta}$ in the definition of $G_n(u)$ will not change the resulting estimator asymptotically. This requires uniform convergence of G_n under some conditions on the sampling design. The required regularity conditions and the related technical details are similar to those used in Appendix 2 of Wu and Sitter (2001b) where the same kind of problem was encountered.

Variance estimation for the proposed MPEML estimators will be addressed in Section 4. A simple and stable algorithm to compute the weights $\hat{p}_i$ is available in Chen, Sitter and Wu (2002).

2.2. Obtain fitted values from a generalized linear model

It is very attractive to directly adapt a generalized linear model as a working model for $g_i = E_\xi(I(y_i \le y_0)|\boldsymbol{x}_i) = P(y_i \le y_0|\boldsymbol{x}_i)$. The extensive literature on binary data analysis can be borrowed here to get fitted values for g_i. For instance, we may use a logistic regression model

$$\log\left(\frac{g_i}{1-g_i}\right) = \boldsymbol{x}_i'\boldsymbol{\theta}\,, \tag{6}$$

with variance function $V(g) = g(1-g)$. Under such a model the finite population parameter $\boldsymbol{\theta}_N$ can be defined as a solution to the optimal estimating equations based on the entire finite population, $\sum_{i=1}^N \boldsymbol{x}_i(z_i^* - g_i) = \mathbf{0}$, where $z_i^* = I(y_i \le y_0)$, $g_i = \exp(\boldsymbol{x}_i'\boldsymbol{\theta})/\{1+\exp(\boldsymbol{x}_i'\boldsymbol{\theta})\}$. A design-based estimator $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}_N$ can be obtained by solving a sample-based version of the system $\sum_{i\in s} d_i\boldsymbol{x}_i(z_i^* - g_i) = \mathbf{0}$. Let $g_i(\boldsymbol{\theta}_N) = \exp(\boldsymbol{x}_i'\boldsymbol{\theta}_N)/\{1+\exp(\boldsymbol{x}_i'\boldsymbol{\theta}_N)\}$. Our second proposed MPEML estimator $\hat{F}_{EL}^{(2)}(y)$ is now similarly defined by using the $g_i(\hat{\boldsymbol{\theta}})$ in the constraint (4). Let $\bar{z}_d$, $\bar{g}_N$, $\bar{g}_d$ and B_N be similarly defined as in Theorem 1, but replace $g_i(\boldsymbol{\theta}_N, \sigma_N, y_0)$ by $g_i(\boldsymbol{\theta}_N)$.

Theorem 2. *The model-calibrated pseudo empirical maximum likelihood estimator $\hat{F}_{EL}^{(2)}(y)$ is asymptotically equivalent to a generalized regression estimator: $\hat{F}_{EL}^{(2)}(y) = \bar{z}_d + (\bar{g}_N - \bar{g}_d)B_N + o_p(n^{-1/2})$. Hence $\hat{F}_{EL}^{(2)}(y)$ is asymptotically design-consistent estimator of $F_Y(y)$. It is also approximately model-unbiased under* (6) *for $y = y_0$.*

Proof of the theorem is similar to that of Theorem 1 and is omitted. One of the advantages of using a generalized linear model is that the error distribution in the regression model is no longer an issue. The logistic regression model gives a reasonable fit in most situations.

2.3. Obtain pseudo fitted values from a semi-parametric model

The variable $z_i = I(y_i \le y)$ only takes values of 0 or 1, but the fitted values g_i from Sections 2.1 and 2.2 are always between 0 and 1. It is tempting to use what we called pseudo fitted values $g_i(\hat{\boldsymbol{\theta}}) = I(\hat{y}_i \le y_0)$ which are also $0-1$ variates, where $\hat{y}_i$ are fitted values for y_i. Under a semi-parametric model $E_\xi(y_i|\boldsymbol{x}_i) = \mu_i$, $V_\xi(y_i|\boldsymbol{x}_i) = v(\mu_i)$, where $\mu_i = \mu(\boldsymbol{x}_i, \boldsymbol{\theta})$ and $v(\cdot)$ is a variance function, and the fitted values $\hat{y}_i$ are given by $\mu(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}})$. Let $h(\cdot)$ be a known link function such that $h(\mu_i) = \boldsymbol{x}_i'\boldsymbol{\theta}$. The $\hat{\boldsymbol{\theta}}$ is the maximum quasi-likelihood estimator of $\boldsymbol{\theta}$ given by solving estimating equations: $\sum_{i\in s}\{d_i\boldsymbol{x}_i(y_i - \mu_i)\}/\{v(\mu_i)h'(\mu_i)\} = \mathbf{0}$, where $h'(u) = dh(u)/du$. Let $\boldsymbol{\theta}_N$ be the solution to $\sum_{i=1}^N\{\boldsymbol{x}_i(y_i - \mu_i)\}/\{v(\mu_i)h'(\mu_i)\} = \mathbf{0}$.

We similarly define the third MPEML estimator $\hat{F}_{EL}^{(3)}(y)$ as before but use the pseudo fitted values $g_i(\hat{\boldsymbol{\theta}}) = I(\hat{y}_i \le y_0)$ in the constraint (4).

Under the same conditions of Theorem 1, it can be shown that $\hat{F}_{EL}^{(3)}(y) = \bar{z}_d + (\bar{g}_N - \bar{g}_d)B_N + o_p(n^{-1/2})$, where $\bar{z}_d$, $\bar{g}_N$, $\bar{g}_d$ and B_N are similarly defined as in Theorem 1 but use $g_i(\boldsymbol{\theta}_N) = I\{\mu(\boldsymbol{x}_i, \boldsymbol{\theta}_N) \le y_0\}$ in place of $g_i(\boldsymbol{\theta}_N, \sigma_N, y_0)$.

It follows that $\hat{F}_{EL}^{(3)}(y)$ is asymptotically design-unbiased estimator for $F_Y(y)$. But $\hat{F}_{EL}^{(3)}(y)$ is not approximately model-unbiased since $E_\xi\{I(y_i \le y)|\boldsymbol{x}\} \ne I\{\mu(\boldsymbol{x}_i, \boldsymbol{\theta}) \le y\}$. However, $\hat{F}_{EL}^{(3)}(y)$ possesses properties not enjoyed by $\hat{F}_{EL}^{(1)}(y)$ and $\hat{F}_{EL}^{(2)}(y)$.

An ad hoc argument for using $\hat{F}_{EL}^{(3)}(y)$ is that if the model fits the finite population perfectly, i.e., $y_i = \mu(\boldsymbol{x}_i, \boldsymbol{\theta}_N)$, $i = 1, \ldots, N$, then $g_i(\hat{\boldsymbol{\theta}}) = I(y_i \le y_0)$ and $\hat{F}_{EL}^{(3)}(y_0)$ reduces to the exact value of $F_Y(y_0)$. It can be expected that in the case of strong auxiliary information, the correlation between y_i and $\hat{y}_i$ is high, and consequently, $\hat{F}_{EL}^{(3)}(y)$ will perform well.

Under a simple linear model with a single x variable, $\mu(\boldsymbol{x}_i, \boldsymbol{\theta}) = \theta_0 + \theta_1 x_i$, and

$$\frac{1}{N}\sum_{i=1}^N g_i(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^N I(\theta_0 + \theta_1 x_i \le y_0) = F_X((y_0 - \theta_0)/\theta_1)\,,$$

where $F_X(u)$ is the finite population distribution function of the x variable. The constraint (4) reduces to

$$\sum_{i\in s} p_i = 1, \quad \sum_{i\in s} p_i I(\hat{\theta}_0 + \hat{\theta}_1 x_i \le y_0) = F_X((y_0 - \hat{\theta}_0)/\hat{\theta}_1) \quad (0 < p_i < 1)\,,$$

so only the frequency distribution of x needs to be known for the implementation of $\hat{F}_{EL}^{(3)}(y)$.

In all three cases, the proposed estimators are asymptotically equivalent to a generalized regression estimator and are themselves distribution functions. The location y_0 can be determined based on efficiency requirement. We will discuss this issue more in Sections 3 and 5.

3. Quantile Estimation

Let $0 < t < 1$. The tth quantile corresponding to a cdf $F(y)$ is $\xi(t) = F^{-1}(t) = \inf\{y : F(y) \geq t\}$. The conventional estimator for $\xi(t)$ of a finite population without using auxiliary information is obtained by inverting the Horvitz-Thompson estimator for the distribution function: $\hat{\xi}_{HT}(t) = \hat{F}_{HT}^{-1}(t)$, where $\hat{F}_{HT}(y) = \sum_{i\in s} d_i I(y_i \leq y)/\sum_{i\in s} d_i$.

3.1. Quantile estimation through direct inversion

An efficient estimator for $\xi(t)$ can be obtained by inverting an efficient estimator for the distribution function. Let $\hat{F}_{EL}(y)$ be any one of the estimators from Section 2, with the choice of $y_0 = \hat{\xi}_{HT}(t)$, then $\hat{F}_{EL}(y)$ will be more efficient than $\hat{F}_{HT}(y)$ for y in the neighborhood of $\xi(t)$. For any $t \in (0,1)$, let $\hat{\xi}_{EL}(t) = \hat{F}_{EL}^{-1}(t)$. Since $\hat{F}_{EL}(y)$ is a distribution function, the above inversion is computationally simple. Without loss of generality, assume $y_1 \leq y_2 \leq \cdots \leq y_n$, it can be seen that $\hat{\xi}_{EL}(t) = y_k$, where k is determined by $\sum_{i=1}^{k} \hat{p}_i \geq t$ and $\sum_{i=1}^{k-1} \hat{p}_i < t$.

To ease presentation, the following sampling designs will be referred to as type I: (1) simple random sampling with or without replacement; (2) stratified simple random sampling with or without replacement; (3) single stage unequal probability sampling with replacement; (4) multi-stage sampling with first stage clusters sampled with replacement. In the case of with-replacement design, the Hansen-Hurwitz type estimator will be used instead of the H-T estimator, i.e., $\pi_i = nq_i$, where n is the number of draws, q_i is the probability of selecting unit i from each of the n draws. Results from draw to draw are independent.

A weak version of Bahadur representation for quantile process $\hat{\xi}_{EL}(t)$ can be established under type I sampling designs. The proof is given in Appendix 2.

Theorem 3. *Under a type I sampling design and conditions specified in Appendix 2,*

$$\hat{\xi}_{EL}(t) - \xi(t) = \{f(\xi(t))\}^{-1}\{t - \hat{F}_{EL}(\xi(t))\} + o_p(n^{-1/2}),$$

where $\hat{F}_{EL}(y)$ is one of the model-calibrated pseudo empirical likelihood estimators from which $\hat{\xi}_{EL}$ is obtained, $f(\cdot)$ is the density function of the limiting distribution function of $F_Y(y)$ as $N \to \infty$.

The improvement in efficiency from using $\hat{\xi}_{EL}(t)$ over $\hat{\xi}_{HT}(t)$ is comparable to that of $\hat{F}_{EL}(y)$ over $\hat{F}_{HT}(y)$. With the optimal choice of $g_i = E_\xi(z_i|\boldsymbol{x}_i)$, the maximum gain of asymptotic efficiency is guaranteed. The major advantage of the model-calibrated pseudo empirical likelihood approach is to achieve this high efficiency while maintaining the computational simplicity. The method can be easily applied to complex sampling designs and multivariate auxiliary variables.

3.2. Quantile estimation using a model-calibrated difference estimator

A model-calibrated difference estimator (MD) for $\xi(t)$ can be derived as follows. Let $\hat{y}_i = \mu(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}})$, $i = 1, \ldots, N$, be the fitted values from a working model. Let

$$\hat{F}_U(y) = \frac{1}{N}\sum_{i=1}^{N} I(\hat{y}_i \le y) \quad \text{and} \quad \hat{F}_V(y) = (\sum_{i\in s} d_i)^{-1} \sum_{i\in s} d_i I(\hat{y}_i \le y)\,.$$

$\hat{F}_U(y)$ can be viewed as a prediction estimator for $F_Y(y)$ while $\hat{F}_V(y)$ is a design-based estimate for $\hat{F}_U(y)$. They are both distribution functions. Note that, as candidate estimators for $F_Y(y)$, neither of them can be better than $\hat{F}_{HT}(y)$. Let $\hat{\xi}_U(t) = \hat{F}_U^{-1}(t)$ and $\hat{\xi}_V(t) = \hat{F}_V^{-1}(t)$. A model-calibrated difference estimator of $\xi(t)$ can then be constructed as

$$\hat{\xi}_{MD}(t) = \hat{\xi}_{HT}(t) - \hat{\xi}_V(t) + \hat{\xi}_U(t)\,. \tag{7}$$

Similar to the construction of $\hat{F}_{EL}^{(3)}(y)$, an ad hoc motivation behind $\hat{\xi}_{MD}(t)$ is that, in the case of no modeling variation for the finite population, i.e., $y_i = \hat{y}_i$, $i = 1, \ldots, N$, we have $\hat{\xi}_U(t) = \xi(t)$, $\hat{\xi}_V(t) = \hat{\xi}_{HT}(t)$, and this model-calibrated difference estimator $\hat{\xi}_{MD}(t)$ reduces to the exact value $\xi(t)$. This estimator is closely related to the difference estimator for the distribution function, $\hat{F}_{MD}(y) = \hat{F}_{HT}(y) - \hat{F}_V(y) + \hat{F}_U(y)$. It has been shown by Rao *et al.* (1990) that $\hat{F}_{MD}$ usually performs better than the Horvitz-Thompson estimator $\hat{F}_{HT}(y)$, and therefore $\hat{\xi}_{MD}(t)$ can also be expected to perform better than $\hat{\xi}_{HT}(t)$. Note that $\hat{F}_{MD}(y)$ itself is not a distribution function, obtaining estimates for quantiles through direct inversion of $\hat{F}_{MD}(y)$ may be difficult or undesirable. The proposed strategy successfully bypasses this difficulty and the resulting estimator is efficient and computationally simple.

Under a linear model with a single x variable, $\hat{y}_i = \hat{\theta}_0 + \hat{\theta}_1 x_i$, $\hat{\xi}_{MD}(t)$ reduces to $\hat{\xi}_{MD}(t) = \hat{\xi}_{HT}(t) + \{\xi_X(t) - \hat{\xi}_X(t)\}\hat{\theta}_1$, where $\xi_X(t)$ and $\hat{\xi}_X(t)$ are the tth population quantile of the x variable and its Horvitz-Thompson estimator, $\hat{\theta}_1$ is the estimated regression coefficient of y over x (assume $\hat{\theta}_1 > 0$).

4. Variance Estimation and Confidence Intervals

4.1. Variance estimation for the distribution function

Let $\hat{F}_{EL}(y)$ be any one of the estimators discussed in Section 2, let $g_i(\boldsymbol{\theta}_N)$ and B_N be defined accordingly. The asymptotic design-based variance of $\hat{F}_{EL}(y)$ is the same as that of $(\sum_{i\in s} d_i)^{-1} \sum_{i\in s} d_i\{I(y_i \le y) - g_i(\boldsymbol{\theta}_N)B_N\}$ which is a ratio type estimator. Let V_p denote the design-based variance.

Theorem 4. *Under conditions specified in Appendix* 1 *and a fixed sample size design, the asymptotic design-based variance of* $\hat{F}_{EL}(y)$ *is*

$$V_p\{\hat{F}_{EL}(y)\} = \frac{1}{N^2}\sum_{i<j}\sum_{j=1}^{N}(\pi_i\pi_j - \pi_{ij})\left(\frac{U_i}{\pi_i} - \frac{U_j}{\pi_j}\right)^2 + o(n^{-1}),$$

where $U_i = I(y_i \le y) - F_Y(y) - \{g_i(\boldsymbol{\theta}_N) - \bar{g}_N\}B_N$, $\bar{g}_N = N^{-1}\sum_{i=1}^N g_i(\boldsymbol{\theta}_N)$. $V_p\{\hat{F}_{EL}(y)\}$ *can be consistently estimated by*

$$v\{\hat{F}_{EL}(y)\} = \frac{1}{N^2}\sum_{i<j}\sum_{j\in s}\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\left(\frac{u_i}{\pi_i} - \frac{u_j}{\pi_j}\right)^2,$$

where $u_i = I(y_i \le y) - \hat{F}_{EL}(y) - \{g_i(\hat{\boldsymbol{\theta}}) - \bar{g}\}\hat{B}$, *and* $\bar{g}$ *and* $\hat{B}$ *are sample-based estimates for* $\bar{g}_N$ *and* B_N.

4.2. Confidence intervals for the distribution function

Let $v = v[\hat{F}(y_0)]$ be the estimated variance of $\hat{F}(y)$ at $y = y_0$, where the estimator $\hat{F}(y)$ for $F(y)$ may be chosen to be (but not necessarily restricted to be) one of those in Section 2. The conventional $1-\alpha$ confidence interval for $F_Y(y_0)$ is

$$(\hat{F}(y_0) - z_{\alpha/2}v^{1/2}, \quad \hat{F}(y_0) + z_{\alpha/2}v^{1/2}), \tag{8}$$

where $z_{\alpha/2}$ is $1-\alpha/2$ quantile from $N(0,1)$. The validity of (8) relies on the asymptotic normality of $\hat{F}(y_0)$,

$$\{\hat{F}(y_0) - F_Y(y_0)\}/v^{1/2} \xrightarrow{L} N(0,1), \tag{9}$$

which can be justified in many situations (see, for example, Francisco and Fuller (1991), Theorem 2). However, for small to moderate sample size, due to the range constraint $0 \le \hat{F}(y) \le 1$, the sampling distribution of $\hat{F}(y)$ for y at large or small quantiles is usually not symmetric. The finite sample performance of confidence intervals (8) at those quantiles is often not satisfactory: the coverage probability is usually lower than the nominal value, and the two tail probabilities are seriously unbalanced (Wu (1999)).

A simple transformation technique can be employed here to construct better behaved confidence intervals for the distribution function.

If (9) holds, then for any monotone smooth function g, $\hat{W} = g(\hat{F}(y_0))$ also has an asymptotic normal distribution with mean $W \doteq g(F_Y(y_0))$ and variance $Var(\hat{W}) \doteq [g'\{F_Y(y_0)\}]^2 \; Var\{\hat{F}(y_0)\}$ (Shao and Tu (1995), p.448), where $g'(u) = dg(u)/du$. Let $v(\hat{W}) = [g'\{\hat{F}(y_0)\}]^2 \; v\{\hat{F}(y_0)\}$. A $1-\alpha$ confidence interval for $F_Y(y_0)$ can be constructed by first obtaining a $1-\alpha$ normal confidence interval for $W = g(F_Y(y_0))$ and then transforming back to $F_Y(y_0)$. This gives the following transformed confidence interval for $F_Y(y_0)$,

$$(g^{-1}\{\hat{W} - z_{\alpha/2}v^{1/2}(\hat{W})\}\,,\ \ g^{-1}\{\hat{W} + z_{\alpha/2}v^{1/2}(\hat{W})\})\,. \tag{10}$$

The transformation is chosen such that the distribution of $\hat{W}$ is better approximated by the normal distribution. Two such transformations are readily available, namely, the logit transformation and the complementary log-log transformation: $\hat{W}_1 = \log\{\hat{F}(y_0)/(1 - \hat{F}(y_0))\}$ and $\hat{W}_2 = \log\{-\log(\hat{F}(y_0))\}$. The resulting confidence intervals using (10) both have simple closed forms:

$$(\hat{F}(y_0)C_{1L},\quad \hat{F}(y_0)C_{1U}) \tag{11}$$

$$(\hat{F}(y_0)^{C_{2L}},\quad \hat{F}(y_0)^{C_{2U}}) \tag{12}$$

where $C_{1L} = \{\hat{F}(y_0) + (1 - \hat{F}(y_0))\exp[z_{\alpha/2}v^{1/2}[\hat{F}(y_0)]/(\hat{F}(y_0)(1 - \hat{F}(y_0)))]\}^{-1}$, $C_{1U} = \{\hat{F}(y_0) + (1 - \hat{F}(y_0))\exp[-z_{\alpha/2}v^{1/2}[\hat{F}(y_0)]/(\hat{F}(y_0)(1 - \hat{F}(y_0)))]\}^{-1}$, $C_{2L} = \exp[-z_{\alpha/2}v^{1/2}[\hat{F}(y_0)]/(\hat{F}(y_0)\log(\hat{F}(y_0)))]$, and $C_{2U} = \exp[z_{\alpha/2}v^{1/2}[\hat{F}(y_0)]/(\hat{F}(y_0)\log(\hat{F}(y_0)))]$.

A simulation study conducted by Wu (1999) shows that (11) and (12) perform like (8) when y_0 is in the middle range of quantiles, but are dramatically superior when y_0 is at small or large quantiles.

5. A Simulation

We conducted a small simulation study to investigate the finite sample performance of the estimators for the distribution function, and the impact of different choices of y_0 on the resulting estimators. More simulation results on the transformed confidence intervals (10) and quantile estimator $\hat{\xi}_{MD}(t)$ can be found in Wu (1999).

Two finite populations of size $N = 2000$ were generated from a regression model $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$, where the x_{1i}'s and x_{2i}'s were generated respectively from a gamma distribution and a lognormal distribution, and the ε_i's are i.i.d. random variates from $N(0, \sigma^2)$. The value of σ^2 is chosen such that for Population 1 the correlation coefficient between y_i and $\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$ is 0.90, and for Population 2 it is 0.70.

At each simulation run, a simple random sample of size $n = 100$ was taken from the finite population and the three model-calibrated pseudo empirical maximum likelihood estimators for $F_Y(y)$ were computed at $y = \xi(t)$ for $t = 0.1, \ldots, 0.9$. The fixed value y_0 used in the calibration equation (4) was preset at $\xi(0.3)$, $\xi(0.5)$ and $\xi(0.7)$ and separate results were obtained under each of these preset values. The process was repeated $B = 1000$ times.

The performance of these estimators was evaluated in terms of Relative Bias (RB) and Relative Efficiency (RE) with $RB = B^{-1}\sum_{b=1}^{B}[\hat{F}_{EL}(y) - F_Y(y)]/F_Y(y)$ and $RE = MSE_{HT}/MSE_{EL}$, where b indexes the bth simulation run, $MSE_{EL} = B^{-1}\sum_{b=1}^{B}[\hat{F}_{EL}(y) - F_Y(y)]^2$, and MSE_{HT} is similarly defined for the baseline Horvitz-Thompson estimator. Table 1 reports the simulated RE for each of the estimators at various population quantiles. The absolute values of the RB's are all less than 1% and are thus not reported.

Table 1. Relative Efficiency of Estimators for $F_Y(y)$ at $y = \xi(t)$ (P1 and P2).

	y_0	$\hat{F}_{EL}(y)$	$t = 0.10$	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90
P1	$\xi(0.3)$	$\hat{F}_{EL}^{(1)}(y)$	1.16	1.32	1.54	1.65	1.71	1.67	1.48	1.29	1.14
		$\hat{F}_{EL}^{(3)}(y)$	1.15	1.24	1.31	1.32	1.33	1.28	1.16	1.08	1.04
	$\xi(0.5)$	$\hat{F}_{EL}^{(1)}(y)$	1.12	1.24	1.47	1.63	1.78	1.89	1.78	1.56	1.27
		$\hat{F}_{EL}^{(2)}(y)$	1.12	1.24	1.46	1.61	1.74	1.82	1.68	1.49	1.26
		$\hat{F}_{EL}^{(3)}(y)$	1.08	1.17	1.35	1.45	1.58	1.52	1.39	1.25	1.13
	$\xi(0.7)$	$\hat{F}_{EL}^{(1)}(y)$	1.05	1.13	1.29	1.43	1.58	1.87	2.11	2.14	1.61
		$\hat{F}_{EL}^{(2)}(y)$	1.05	1.12	1.27	1.41	1.54	1.79	2.00	2.08	1.60
		$\hat{F}_{EL}^{(3)}(y)$	1.03	1.07	1.16	1.24	1.32	1.50	1.68	1.71	1.32
P2	$\xi(0.3)$	$\hat{F}_{EL}^{(1)}(y)$	1.06	1.08	1.11	1.19	1.21	1.24	1.29	1.22	1.18
		$\hat{F}_{EL}^{(2)}(y)$	1.06	1.08	1.10	1.17	1.19	1.22	1.27	1.20	1.16
		$\hat{F}_{EL}^{(3)}(y)$	1.02	1.02	1.01	1.04	1.03	1.01	1.02	1.00	1.00
	$\xi(0.5)$	$\hat{F}_{EL}^{(1)}(y)$	1.05	1.08	1.10	1.18	1.21	1.27	1.34	1.32	1.28
		$\hat{F}_{EL}^{(2)}(y)$	1.05	1.07	1.09	1.16	1.19	1.25	1.32	1.29	1.26
		$\hat{F}_{EL}^{(3)}(y)$	1.04	1.06	1.07	1.11	1.11	1.14	1.16	1.12	1.09
	$\xi(0.7)$	$\hat{F}_{EL}^{(1)}(y)$	1.04	1.06	1.09	1.15	1.20	1.27	1.38	1.42	1.43
		$\hat{F}_{EL}^{(2)}(y)$	1.04	1.06	1.09	1.14	1.18	1.25	1.34	1.39	1.40
		$\hat{F}_{EL}^{(3)}(y)$	1.01	1.03	1.05	1.08	1.12	1.17	1.22	1.28	1.21

Table 1 can be summarized as follows: (i) $\hat{F}_{EL}^{(1)}$ and $\hat{F}_{EL}^{(2)}$ have very similar performance and are better than $\hat{F}_{EL}^{(3)}$ in all cases; (ii) the most efficient estimator for $F_Y(y)$ is obtained by setting $y_0 = y$ in (4), but the estimator $\hat{F}_{EL}(y)$ with a prespecified value y_0 is also very efficient when y is close to y_0, though it becomes less efficient when y is far away from y_0; (iii) with strong auxiliary information (Population 1), the gain from using $\hat{F}_{EL}(y)$ can be substantial compared to the baseline Horvitz-Thompson estimator, and when auxiliary information becomes weak or less relevant (Population 2), the superiority of our proposed estimators gradually disappears though they are never worse than the HT estimator; (iv) for the two populations considered here, higher efficiency occurs at higher population quantiles (large y) (this last point due to the skewed distribution of Y, the $\boldsymbol{X}$ variables are better predictors for the response variable Y at the higher quantiles).

We also considered other finite populations where the regression model is misspecified. For instance, we generated finite populations from a superpopulation model $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_{11} x_{1i}^2 + \beta_{22} x_{2i}^2 + \beta_{12} x_{1i} x_{2i} + \varepsilon_i$, and then blindly fit a simple linear regression model or a logistic regression model as we did for Populations 1 and 2 for each of the sample data. The conclusions are very much the same as before. The gain of efficiency from using $\hat{F}_{EL}(y)$ depends on the correlation between the $\boldsymbol{X}$ variables and the response variable Y. Knowing the "correct model" will of course maximize the efficiency of those estimators, but any "reasonable" working model will guarantee the gain of efficiency from using the proposed estimators.

6. Concluding Remarks

Estimation of the finite population distribution function and quantiles in the presence of auxiliary information requires special treatments in terms of both modeling and construction. The generalized regression estimator which is the most popular one for the estimation of means and totals cannot be directly applied here to get an estimator with desirable properties. The model-calibrated pseudo empirical maximum likelihood estimators proposed in this article are optimal under a chosen model, very easy to compute and highly efficient with strong auxiliary information. They are also robust against model misspecifications. Implementation of these estimators requires complete auxiliary information in general, i.e., values of the $\boldsymbol{X}$ variables for the entire finite population. When such information is not available, the technique can be used under two-phase sampling where a large first phase sample serves as "complete" auxiliary information. Estimates for quantiles can be obtained directly from inverting the estimated distribution function.

Appendix 1: Proof of Theorem 1.

We assume there is a sequence of finite populations $\{U_\nu,\ \nu = 1, 2, \ldots\}$. $F_\nu(y)$, $\xi_\nu(t)$ refer to $F_Y(y)$ and $\xi(t)$ for the finite population U_ν. The index ν will be suppressed when there is no confusion. All limiting processes are under $\nu \to \infty$.

Let $\mu'(\boldsymbol{x}, \boldsymbol{\theta}) = \partial\mu(\boldsymbol{x}, \boldsymbol{\theta})/\partial\boldsymbol{\theta}$. The following regularity conditions are required.

A1.1 $|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_N| = O_p(n^{-1/2})$ and $\hat{\sigma} - \sigma_N = O_p(n^{-1/2})$.

A1.2 $\max_{i=1,\ldots,N}(nd_i/N) = O(1)$.

A1.3 For each $\boldsymbol{x}_i$, $\mu(\boldsymbol{x}_i, \boldsymbol{\theta})$ is twice differentiable, $N^{-1}\sum_{i=1}^{N}\{\mu'(\boldsymbol{x}_i, \boldsymbol{\theta}_N)/v_i\}^2 =$

$O(1)$, $N^{-1}\sum_{i=1}^N(\mu(\boldsymbol{x}_i,\boldsymbol{\theta}_N)/v_i)^2 = O(1)$, and $N^{-1}\sum_{i=1}^N v_i^{-2} = O(1)$.

Proof of part (1). Suppose that $\boldsymbol{\theta}_N$ and σ_N are known. We use $g_i = g_i(\boldsymbol{\theta}_N, \sigma_N, y_0)$ in (4) to construct $\hat{F}_{EL}^{(1)}(y)$. Let $u_i = g_i(\boldsymbol{\theta}_N, \sigma_N, y_0) - \bar{g}_N$. Since $|u_i| \le 1$, it follows from the proof of Theorem 1 in Chen and Sitter (1999) that $\hat{p}_i = w_i/(1+\lambda u_i)$, $\lambda = \{\sum_{i\in s} w_i u_i\}/\{\sum_{i\in s} w_i u_i^2\} + o_p(n^{-1/2})$, with $w_i = d_i/\sum_{i\in s} d_i$. Note that $\hat{p}_i = w_i(1-\lambda u_i) + o_p(n^{-1/2})$. Hence (5) follows from above expansion.

When $g_i = g_i(\boldsymbol{\theta}_N, \sigma_N, y_0)$ is replaced by $\hat{g}_i = g_i(\hat{\boldsymbol{\theta}}, \hat{\sigma}, y_0)$, we need only show that

$$\frac{1}{N}\sum_{i=1}^N \hat{g}_i - (\sum_{i\in s} d_i)^{-1}\sum_{i\in s} d_i\hat{g}_i = \frac{1}{N}\sum_{i=1}^N g_i - (\sum_{i\in s} d_i)^{-1}\sum_{i\in s} d_i g_i + o_p(n^{-1/2})\,. \quad (13)$$

Condition **A1.2** implies that $N^{-1}\sum_{i\in s} d_i c_i - N^{-1}\sum_{i=1}^N c_i = O_p(n^{-1/2})$ if $N^{-1}\sum_{i=1}^N c_i^2 = O(1)$. Let $a_i(\boldsymbol{\theta},\sigma) = (\partial/\partial\boldsymbol{\theta})g_i(\boldsymbol{\theta},\sigma)$, $b_i(\boldsymbol{\theta},\sigma) = (\partial/\partial\sigma)g_i(\boldsymbol{\theta},\sigma)$. It follows from a Taylor series expansion that

$$\hat{g}_i = g_i + [a_i(\boldsymbol{\theta}_N,\sigma_N)]'(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_N) + b_i(\boldsymbol{\theta}_N,\sigma_N)(\hat{\sigma}-\sigma_N) + O_p(|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_N|^2) + O_p((\hat{\sigma}-\sigma_N)^2)\,.$$

Under the regularity conditions, $N^{-1}\sum_{i=1}^N a_i - (\sum_{i\in s} d_i)^{-1}\sum_{i\in s} d_i a_i = O_p(n^{-1/2})$, $N^{-1}\sum_{i=1}^N b_i - (\sum_{i\in s} d_i)^{-1}\sum_{i\in s} d_i b_i = O_p(n^{-1/2})$. Under condition **A1.1**, (13) follows immediately from the foregoing expansion. This proves the first part of Theorem 1.

Proof of part (2). From part (1), $\hat{F}_{EL}^{(1)}(y) = \hat{F}_{HT}(y) + O_p(n^{-1/2})$, so $\hat{F}_{EL}^{(1)}(y)$ is asymptotically design-consistent. Assume $(\boldsymbol{x}_i, y_i)$, $i = 1,\ldots,N$ are i.i.d. random variates from the superpopulation, then $\theta_N = \theta + O_p(n^{-1/2})$. To prove that $\hat{F}_{EL}^{(1)}(y)$ is approximately model-unbiased at $y = y_0$, we substitute $\hat{g}_i = g_i(\hat{\boldsymbol{\theta}}, \hat{\sigma}, y_0)$ used in (4) by $g_i^* = g_i(\boldsymbol{\theta}, \sigma, y_0)$, denote the resulting estimator by $\hat{F}_{EL}^*(y) = \sum_{i\in s} p_i^* I(y_i \le y)$. Under model (3), we have

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

Hence $\hat{F}_{EL}^*(y_0)$ 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 $\hat{\theta} = \theta_N + O_p(n^{-1/2})$ under the design-based framework and $\theta_N = \theta + O_p(n^{-1/2})$ under the superpopulation model. To conclude that $\hat{\theta} = \theta + O_p(n^{-1/2})$ we need to consider the joint expectation under both the design and the model. Suppose that $\hat{\theta} = \theta + O_p(n^{-1/2})$ under the model, it is then straightforward to show that $\hat{g}_i = g_i^* + o_p(1)$, with the uniform term $o_p(1)$ over i. Similarly, $u_i = u_i^* + o_p(1)$, uniformly over i, where $u_i = \hat{g}_i - N^{-1}\sum_{i=1}^N \hat{g}_i$ and $u_i^* = g_i^* - N^{-1}\sum_{i=1}^N g_i^*$. It now follows that

$\hat{p}_i = w_i/(1+\lambda u_i) + o_p(n^{-1/2}) = w_i/(1+\lambda u_i^*) + o_p(1) = p_i^* + o_p(1)$. The approximate model unbiasedness follows from $\hat{F}_{EL}^{(1)}(y_0) = \hat{F}_{EL}^{*}(y_0) + o_p(1)$.

Appendix 2: Proof of Theorem 3.

To simplify notation, for a fixed $t \in (0,1)$, we use ξ_ν, ξ and $\hat{\xi}$ instead of $\xi_\nu(t)$, $\xi(t)$ and $\hat{\xi}(t)$, etc. The conditions needed are those used in Theorem 1, and **A2.1** and **A2.2** below.

A2.1 There exists a cdf $F(y)$ which is twice differentiable, with density function $f(y)$, such that $F_\nu(y) - F(y) = o(1)$; and for any $a_\nu = O(n^{-1/2})$,

$$\sup_{|\delta|\le a_\nu} |[F_\nu(y+\delta) - F_\nu(y)] - [F(y+\delta) - F(y)]| = o(n_\nu^{-1/2}),$$

where the sample size $n_\nu \to \infty$ as $\nu \to \infty$.

A2.2 For fixed $t \in (0,1)$, $\xi_\nu \to \xi_0$, where ξ_0 is the tth quantile from $F(y)$ and $f(\xi_0) > 0$.

It is straightforward to show that $\hat{\xi}_{EL} - \xi = O_p(n^{-1/2})$. Following Serfling (1980), we need only to show that, for $c_n = O_p(n^{-1/2})$,

$$\sup_{|\delta|\le c_n} |[\hat{F}_{EL}(\xi+\delta) - \hat{F}_{EL}(\xi)] - [F_Y(\xi+\delta) - F_Y(\xi)]| = o_p(n^{-1/2}). \tag{14}$$

We prove (14) in three steps.

Step 1. Let $\hat{G}_{HT}(y) = N^{-1}\sum_{i\in s} d_i I(y_i \le y)$ be the conventional Horvitz-Thompson estimator for $F_Y(y)$. We first investigate the validity of

$$\sup_{|\delta|\le c_n} |[\hat{G}_{HT}(\xi+\delta) - \hat{G}_{HT}(\xi)] - [F_Y(\xi+\delta) - F_Y(\xi)]| = o_p(n^{-1/2}). \tag{15}$$

Shao and Rao (1993) show that (15) is true under stratified multi-stage sampling with first stage sampling of clusters with replacement (see also Chen, Rao and Sitter (2000)). The crucial argument that leads to this result is the Bernstein's inequality (Serfling (1980), p.95). The inequality is valid for simple random sampling with or without replacement (Shorack and Wellner (1986), p.878), and under unequal probability sampling with replacement.

Step 2. Let $\hat{F}_{HT}(y) = \sum_{i\in s} d_i I(y_i \le y)/\sum_{i\in s} d_i$ be the modified Horvitz-Thompson estimator for $F_Y(y)$. Note that $\hat{F}_{HT}(y)$ is a ratio estimator. Condition **A1.3** implies $N^{-1}\sum_{i\in s} d_i = 1 + O_p(n^{-1/2})$. It now follows that

$$\sup_{|\delta|\le c_n} |[\hat{G}_{HT}(\xi+\delta) - \hat{G}_{HT}(\xi)] - [\hat{F}_{HT}(\xi+\delta) - \hat{F}_{HT}(\xi)]| = o_p(n^{-1/2}), \tag{16}$$

since $[\hat{G}_{HT}(\xi+\delta) - \hat{G}_{HT}(\xi)] - [\hat{F}_{HT}(\xi+\delta) - \hat{F}_{HT}(\xi)] = (N^{-1}\sum_{i\in s} d_i - 1)[\hat{F}_{HT}(\xi+\delta) - \hat{F}_{HT}(\xi)]$.

Step 3. The final step is to show that

$$\sup_{|\delta|\le c_n} |[\hat{F}_{EL}(\xi+\delta)-\hat{F}_{EL}(\xi)]-[\hat{F}_{HT}(\xi+\delta)-\hat{F}_{HT}(\xi)]| = o_p(n^{-1/2}), \qquad (17)$$

then the conclusion (14) follows from (15), (16) and (17). Note that $\hat{p}_i = w_i/(1+\lambda u_i) = w_i - w_i\lambda u_i/(1+\lambda u_i)$, where $w_i = d_i/\sum_{i\in s} d_i$. It is straightforward to see that, for any $s_1 < s_2$,

$$[\hat{F}_{EL}(s_2)-\hat{F}_{EL}(s_1)]-[\hat{F}_{HT}(s_2)-\hat{F}_{HT}(s_1)] = -\lambda\sum_{i\in s}\frac{w_iu_i}{1+\lambda u_i}I(s_1<y_i\le s_2).$$

Since $|u_i|\le 1$, (17) follows from the argument of Lemma 2 of Chen and Chen (2000).

Acknowledgements

This research was supported by grants from the Natural Sciences and Engineering Research Council of Canada.

References

Chambers, R. L. and Dunstan, R. (1986). Estimating distribution function from survey data. *Biometrika* **73**, 597-604.

Chen, H. and Chen, J. (2000). Bahadur representations of the empirical likelihood quantile processes. *Nonparametric Statist.* **12**, 645-660.

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

Chen, J. and Sitter, R. R. (1999). A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statist. Sinica* **9**, 385-406.

Chen, J., Rao, J. N. K. and Sitter, R. R. (2000). Efficient random imputation for missing data in complex surveys. *Statist. Sinica* **10**, 1153-1169.

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

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

Francisco, C. A. and Fuller, W. A. (1991). Quantiles estimation with a complex survey design. *Ann. Statist.* **19**, 454-469.

Godambe, V. P. and Thompson, M. E. (1986). Parameters of superpopulation and survey population: their relationships and estimation. *Internat. Statist. Rev.* **54**, 127-138.

Rao, J. N. K., Kovar, J. G. and Mantel, H. J. (1990). On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika* **77**, 365-375.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics.* John Wiley, New York.

Shao, J. and Rao, J. N. K. (1993). Standard errors for low income proportions estimated from stratified multi-stage samples. *Sankhyā* B **55**, 393-414.

Shao, J. and Tu, D. (1995). *The Jackknife and Bootstrap.* Springer-Verlag, New York.

Shorack, G. R. and Wellner, J. A. (1986). *Empirical Processes with Applications to Statistics.* Wiley, New York.

Wang, S. and Dorfman, A. H. (1996). A new estimator for the finite population distribution function. *Biometrika* **83**, 639-652.

Wu, C. (1999). The effective use of complete auxiliary information from survey data. Unpublished doctoral dissertation, Simon Fraser University, Canada.

Wu, C. (2001). A note on the optimality of the model calibration estimator. Working paper 2001-03, Department of Statistics and Actuarial Science, University of Waterloo.

Wu, C. and Sitter, R. R. (2001a). A model-calibration approach to using complete auxiliary information from survey data. *J. Amer. Statist. Assoc.* **96**, 185-193.

Wu, C. and Sitter, R. R. (2001b). Variance Estimation for the Finite Population Distribution Function with Complete Auxiliary Information. *Canad. J. Statist.* **29**, 289-307.

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

E-mail: jhchen@math.uwaterloo.ca

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

E-mail: cbwu@math.uwaterloo.ca

(Received May 2001; accepted May 2002)