Journal of Official Statistics, Vol. 19, No. 2, 2003, pp. 119–131

Optimal Calibration Estimators Under Two-Phase Sampling

Changbao Wu[1] *and Ying Luan*[2]

Optimal calibration estimators require in general complete auxiliary information. When such information is not available, the estimation procedure can be combined with two-phase sampling where a large, less costly first-phase sample measured over the auxiliary variables is used to get estimates for certain population quantities related to the covariates. In this article we propose optimal calibration estimators for the population mean, the distribution function, the population variance and other second-order finite population quantities under two-phase sampling. The proposed optimal calibration estimators for a second-order finite population quantity such as the population variance can ideally be used to obtain more efficient variance estimators for a first-order finite population quantity such as the total or the distribution function. The estimation strategy can be applied to various measurement error and non-response problems. The design-based finite sample performances of proposed estimators are investigated through a simulation study using real survey data from the 1996 Statistics Canada Family Expenditure (FAMEX) Survey.

Key words: Auxiliary information; measurement error; model-assisted approach; nonresponse; variance estimation.

1. Introduction

Many highly efficient estimation techniques in survey sampling require strong information about auxiliary variables, $\boldsymbol{x}$. For example, the calibration estimator of Deville and Särndal (1992) requires $\boldsymbol{X}$, the population totals. When such information is not available, a two-phase sampling scheme can be used where a large, less costly first-phase sample measured over the $\boldsymbol{x}$ variable is used to obtain a good estimate of $\boldsymbol{X}$. A smaller and more costly second-phase sample can then be taken and the study variable y is observed. The major advantage of using two-phase sampling is the gain in high precision without substantial increase in cost. Särndal et al. (1992, Chapter 9) provide an excellent account of two-phase sampling.

Recently, Wu and Sitter (2001) proposed a model-calibration approach to using auxiliary information from surveys. The proposed model-calibration estimator is not only highly efficient but also optimal among a class of calibration estimators (Wu

[1] Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON, N2L 3G1, Canada. Email: cbwu@uwaterloo.ca
[2] Department of Statistics, University of California, Riverside, CA 92521, USA.
Acknowledgments: This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada. The authors thank Professor Randy R. Sitter for helpful comments. Thanks are also due to the Associate Editor and three anonymous referees for constructive suggestions and comments that greatly improved the presentation.

2002). To compute the model-calibration estimator, however, one generally requires the values of the $\boldsymbol{x}$ variables to be known for the entire finite population. In practice this complete auxiliary information is often unavailable. Two-phase sampling provides an ideal solution for the use of optimal calibration estimators under such situations.

Suppose $U = \{1, 2, \ldots, N\}$ is the set of labels for the finite population and s is the set of labels for the sampled units. Let y_i and $\boldsymbol{x}_i$ be the values of the response variable y and the vector of covariates $\boldsymbol{x}$ associated with the ith unit. Let π_i be the first-order inclusion probabilities. Assuming the population totals $\boldsymbol{X} = \sum_{i=1}^{N} \boldsymbol{x}_i$ are known, the conventional calibration estimator (Deville and Särndal 1992) for the finite population total $Y = \sum_{i=1}^{N} y_i$ is constructed as $\hat{Y}_C = \sum_{i \in s} w_i y_i$, where the weights w_i minimize a distance measure Φ_s between the w_i's and the basic design weights $d_i = 1/\pi_i$ subject to the so-called benchmark constraints

$$\sum_{i \in s} w_i \boldsymbol{x}_i = \boldsymbol{X} \tag{1}$$

The $\boldsymbol{x}_i$ used in (1) are also referred to as calibration variables.

There are two basic components in the construction of a calibration estimator: a distance measure Φ_s and a set of constraints such as (1). The chi-squared distance measure $\Phi_s = \sum_{i \in s} (w_i - d_i)^2/(d_i q_i)$ is commonly used, where the weighting factors q_i are unrelated to d_i. Other distance measures can also be considered but it has been shown by Deville and Särndal (1992) that the resulting calibration estimator is asymptotically equivalent to the one using a chi-squared distance measure with a certain choice of q_i. The benchmark constraints (1) which calibrate directly over the individual $\boldsymbol{x}$ variables are indeed ad hoc. There is no compelling reason to exclude the use of other types of constraints.

Optimal calibration estimators have been studied by Wu (2002) under a model-assisted framework. The following semi-parametric superpopulation model was used to motivate the optimality considerations,

$$E_\xi(y_i \mid \boldsymbol{x}_i) = \mu(\boldsymbol{x}_i, \boldsymbol{\theta}), \quad V_\xi(y_i \mid \boldsymbol{x}_i) = [v(\boldsymbol{x}_i)]^2 \sigma^2 \tag{2}$$

where $\mu(\cdot, \cdot)$ and $v(\cdot)$ are known functions, $\boldsymbol{\theta}$ and σ^2 are unknown model parameters, and E_ξ, V_ξ denote the expectation and the variance under the model, ξ. It was also assumed that $y_1, y_2, \ldots, y_N$ are conditionally independent given the $\boldsymbol{x}_i$'s. Wu (2002) considered the class of calibration estimators for the population total Y obtained by using

(a) any chi-squared distance measure with the weighting factors satisfying $q_i > q$ for some constant $q > 0$ and $N^{-1} \sum_{i=1}^{N} q_i^2 = O(1)$;
(b) any constraint through a dimension-reduction variable $u_i = u(\boldsymbol{x}_i, \boldsymbol{\theta})$, i.e.,

$$\sum_{i \in s} w_i u(\boldsymbol{x}_i, \boldsymbol{\theta}) = \sum_{i=1}^{N} u(\boldsymbol{x}_i, \boldsymbol{\theta}) \tag{3}$$

where the functional form of $u(\cdot, \cdot)$ can be arbitrary.

Some of the major results can be summarized as follows:

(i) For any consistent estimator of $\boldsymbol{\theta}$ such that $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta} + o_p(1)$, replacing $\boldsymbol{\theta}$ by $\hat{\boldsymbol{\theta}}$ in the constraint (3) will not change the resulting calibration estimator asymptotically.
(ii) Let $\hat{\boldsymbol{\theta}} = (\sum_{i \in s} d_i q_i \boldsymbol{x}_i \boldsymbol{x}_i^T)^{-1} \sum_{i \in s} d_i q_i \boldsymbol{x}_i y_i$. If we use $u_i = \boldsymbol{x}_i^T \hat{\boldsymbol{\theta}}$ as calibration

variable in (3), where T denotes transpose, the resulting calibration estimator is identical to the conventional calibration estimator $\hat{Y}_C$ using (1). Hence, the class of calibration estimators considered by Wu (2002) is very general and includes the conventional calibration estimator as a special member.

(iii) The model-calibration estimator $\hat{\hat{Y}}_{MC}$ of Wu and Sitter (2001) obtained by using $u_i = E_\xi(y_i \,|\, \boldsymbol{x}_i) = \mu(\boldsymbol{x}_i, \boldsymbol{\theta})$ in (3) is optimal under the criterion of minimum model expectation of the asymptotic design-based variance, $E_\xi\{AV_p(\hat{Y})\}$. Here AV_p refers to the asymptotic variance under the sampling design, p, and $\hat{Y}$ is any calibration estimator from the class.

(iv) For the estimation of the population distribution function

$$F_Y(t) = N^{-1} \sum_{i=1}^{N} I(y_i \le t)$$

where $I(\cdot)$ denotes the indicator function, the optimal calibration variable is given by $g(\boldsymbol{x}_i, t) = E_\xi[I(y_i \le t)|\boldsymbol{x}_i] = P(y_i \le t \,|\, \boldsymbol{x}_i)$, which is dependent on the particular value of t.

(v) To estimate a second-order finite population quantity such as the population variance or more generally $Q = \sum_{i=1}^{N} \sum_{j=i+1}^{N} \phi(y_i, y_j)$, the optimal calibration variable is $u_{ij} = E_\xi[\phi(y_i, y_j) \,|\, \boldsymbol{x}_i, \boldsymbol{x}_j]$.

For proofs, the required regularity conditions and more discussion we refer to Wu (2002) which is available online at www.stats.uwaterloo.ca/~cbwu/paper.html.

It is clear that optimal calibration estimation requires in general the complete auxiliary information $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$. In this article we focus on the use of optimal calibration estimators under two-phase sampling assuming that the complete auxiliary information is not available but a large first-phase sample on the $\boldsymbol{x}$ variables can be easily collected with affordable cost.

Estimation of the population mean or total using a regression-type estimator under two-phase sampling has been discussed in the literature. Less attention, however, has been given to the estimation of the distribution function, the variance or other second-order finite population quantities under the context of two-phase sampling. In Section 2, we present optimal calibration estimators for various first- and second-order finite population quantities under two-phase sampling under a unified framework. In Section 3, the optimal calibration estimators for second-order finite population quantities are naturally used to construct more efficient variance estimators for the regression estimator for the population mean. Variance estimation for the distribution function is addressed in Section 4. Some empirical results regarding the finite sample design-based performances of the proposed estimators are reported in Section 5 using real survey data of the 1996 Statistics Canada Family Expenditure (FAMEX) Survey for the province of Ontario. Applications to measurement error and nonresponse problems are briefly discussed in Section 6. Some concluding remarks are given in Section 7.

2. Optimal Calibration Estimators Under Two-Phase Sampling

For simplicity of presentation, we consider cases where a relatively large first-phase sample s' of fixed size n' is selected using simple random sampling without replacement,

and $\boldsymbol{x}_i$ is measured for all $i \in s'$. A second-phase sample s of size n is selected using a general sampling design $p \mid s'$ with first- and second-order inclusion probabilities $\pi_{i|s'}$ and $\pi_{ij|s'}$, and the response variable y_i is observed for $i \in s$. The estimators presented below, however, can be extended to cases where the first-phase sampling design is also arbitrary. Let $d_{i|s'} = 1/\pi_{i|s'}$ and $d_i = (N/n')d_{i|s'}$.

2.1. Estimating the population total

Under Model (2), the model-calibration estimator for the population total Y is defined as $\hat{Y}_{MC} = \sum_{i \in s} w_i y_i$, where the calibrated weights w_i minimize

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

subject to

$$\sum_{i \in s} w_i \mu(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}) = (N/n') \sum_{i \in s'} \mu(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}) \tag{4}$$

Note that the right-hand side of (4) is an estimate for $\sum_{i=1}^{N} \mu(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}})$ based on the first-phase sample s'. It is straightforward to show that

$$\hat{Y}_{MC} = \frac{N}{n'} \left\{ \sum_{i \in s} d_{i|s'} y_i + \left[\sum_{i \in s'} \mu(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}) - \sum_{i \in s} d_{i|s'} \mu(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}) \right] \hat{B}_Y \right\}$$

where $\hat{B}_Y = \sum_{i \in s} d_{i|s'} q_i \hat{\mu}_i y_i / \sum_{i \in s} d_{i|s'} q_i \hat{\mu}_i^2$ and $\hat{\mu}_i = \mu(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}})$.

It can be shown that, under some mild regularity conditions, $\hat{Y}_{MC}$ is asymptotically equivalent to

$$Y^*_{MC} = \frac{N}{n'} \left\{ \sum_{i \in s} d_{i|s'} y_i + \left[\sum_{i \in s'} \mu(\boldsymbol{x}_i, \boldsymbol{\theta}_N) - \sum_{i \in s} d_{i|s'} \mu(\boldsymbol{x}_i, \boldsymbol{\theta}_N) \right] B_Y \right\}$$

where $B_Y = \sum_{i=1}^{N} q_i \mu_i y_i / \sum_{i=1}^{N} q_i \mu_i^2$, $\mu_i = \mu(\boldsymbol{x}_i, \boldsymbol{\theta}_N)$, and $\boldsymbol{\theta}_N$ are the finite population parameters estimated by $\hat{\boldsymbol{\theta}}$. See Wu and Sitter (2001) for a detailed discussion on the estimation of model parameters. The required regularity conditions were detailed in an unpublished master's essay at the University of Waterloo (Luan 2001). It follows that

$$AV_p(\hat{Y}_{MC}) = V_p(Y^*_{MC}) = N^2 \left(\frac{1}{n'} - \frac{1}{N} \right) S_Y^2 + \left(\frac{N}{n'} \right)^2 E_1 V_2 \left[\sum_{i \in s} d_{i|s'} (y_i - B_Y \mu_i) \right]$$

where E_1 and V_2 denote the expectation and the variance under the first-phase and the second-phase sampling design, respectively, and S_Y^2 is the finite population variance for the y variable. Since the value of S_Y^2 is unaffected by different choices of the calibration variable u_i and $E_\xi E_1 V_2[\cdot] = E_1\{E_\xi V_2\}[\cdot]$, following the same argument as in Theorem 1 of Wu (2002), we can show that $\hat{Y}_{MC}$ minimizes $E_\xi[AV_p(\hat{Y})]$ among the class of calibration estimators $\hat{Y}$ similarly defined as in Section 1 with (3) modified for two-phase sampling in the form of (4).

2.2. Estimating the distribution function

The model-calibrated pseudo-empirical maximum likelihood estimator (ME), which is asymptotically equivalent to the optimal calibration estimator (Wu and Sitter 2001), is

particularly appealing for the estimation of the distribution function. Under two-phase sampling, the ME estimator for $F_Y(t)$ is defined as $\hat{F}_{ME}(t) = \sum_{i \in s} \hat{p}_i I(y_i \leq t)$ where the weights $\hat{p}_i$ maximize the pseudo-empirical log-likelihood function

$$\hat{l}(\boldsymbol{p}) = \frac{N}{n'} \sum_{i \in s} d_{i|s'} \log(p_i) \tag{5}$$

subject to

$$\sum_{i \in s} p_i = 1 \ (0 < p_i < 1) \text{ and } \sum_{i \in s} p_i g(\boldsymbol{x}_i, t) = \frac{1}{n'} \sum_{i \in s'} g(\boldsymbol{x}_i, t) \tag{6}$$

where $g(\boldsymbol{x}_i, t) = E_\xi[I(y_i \leq t)|\boldsymbol{x}_i] = P(y_i \leq t|\boldsymbol{x}_i)$. Under a regression working model

$$y_i = \mu(\boldsymbol{x}_i, \boldsymbol{\theta}) + v(\boldsymbol{x}_i)\varepsilon_i, \quad i = 1, 2, \ldots, N \tag{7}$$

where the ε_i's are independent and identically distributed (iid) as $N(0, \sigma^2)$, we have

$$g(\boldsymbol{x}_i, t) = G[\{y - \mu(\boldsymbol{x}_i, \boldsymbol{\theta})\}/\{v(\boldsymbol{x}_i)\sigma\}]$$

The $G(\cdot)$ is the cumulative distribution function of $N(0, 1)$. In applications the unknown model parameters $\boldsymbol{\theta}$ and σ^2 will have to be replaced by sample-based estimates. Note that if a prespecified $t = t_0$ is used in (6) while the resulting weights $\hat{p}_i$ are used in $\hat{F}_{ME}(t)$ for an arbitrary t, $\hat{F}_{ME}(t)$ will itself be a genuine distribution function. The optimality of $\hat{F}_{ME}(t)$ at $t = t_0$ can be established along the lines of Theorem 2 in Wu (2002). When the regression model (7) is not desirable, a slightly less efficient but more robust logistic regression model can be used to obtain $g(\boldsymbol{x}_i, t)$. See Chen and Wu (2002) or Wu (2002) for details. A simple algorithm for computing the pseudo-empirical likelihood weights $\hat{p}_i$ can be found in Chen et al. (2002).

2.3. Estimating second-order finite population quantities

The population variance, the variance of a linear estimator or more generally a second-order finite population quantity of the form $Q = \sum_{i=1}^{N} \sum_{j=i+1}^{N} \phi(y_i, y_j)$ can also be efficiently estimated through optimal calibration. Under two-phase sampling, the optimal model-calibration estimator for Q is given by $\hat{Q}_{MC} = \sum_{i \in s} \sum_{j>i} w_{ij} \phi(y_i, y_j)$, where the weights w_{ij} minimize the modified distance measure $\Phi_{s^*} = \sum_{i \in s} \sum_{j>i} (w_{ij} - d_{ij})^2/(d_{ij} q_{ij})$ subject to

$$\sum_{i \in s} \sum_{j>i} w_{ij} E_\xi[\phi(y_i, y_j) \,|\, \boldsymbol{x}_i, \boldsymbol{x}_j] = \frac{N(N-1)}{n'(n'-1)} \sum_{i \in s'} \sum_{j>i} E_\xi[\phi(y_i, y_j) \,|\, \boldsymbol{x}_i, \boldsymbol{x}_j] \tag{8}$$

Here $d_{ij} = d_{ij|s'}[N(N-1)]/[n'(n'-1)]$ and $d_{ij|s'} = 1/\pi_{ij|s'}$. The weighting factors q_{ij} in the distance measure are prespecified and a common choice would be $q_{ij} = 1$. This optimal model-calibration estimator $\hat{Q}_{MC}$ can be explicitly expressed as a regression-type estimator, as detailed below for the population variance S_Y^2.

Consider the finite population variance $S_Y^2 = (N-1)^{-1} \sum_{i=1}^{N} (y_i - \overline{Y})^2$ which can be alternatively expressed as $S_Y^2 = [N(N-1)]^{-1} \sum_{i=1}^{N} \sum_{j=i+1}^{N} (y_i - y_j)^2$. Under a linear regression working model

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\theta} + v(\boldsymbol{x}_i)\varepsilon_i, \quad i = 1, 2, \ldots, N \tag{9}$$

where the ε_i's are iid with mean zero and variance σ^2, we have

$$E_\xi[(y_i - y_j)^2 \mid \boldsymbol{x}_i, \boldsymbol{x}_j] = \boldsymbol{\theta}^T(\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^T\boldsymbol{\theta} + \sigma^2[v^2(\boldsymbol{x}_i) + v^2(\boldsymbol{x}_j)]$$

Let $u_{ij} = \hat{\boldsymbol{\theta}}^T(\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^T\hat{\boldsymbol{\theta}} + \hat{\sigma}^2[v^2(\boldsymbol{x}_i) + v^2(\boldsymbol{x}_j)]$ and $v_{ij} = (y_i - y_j)^2$. It can be shown that the resulting optimal model-calibration estimator for S_Y^2 is given by

$$\hat{S}_{MC}^2 = \frac{1}{n'(n'-1)}\left[\sum_{i\in s}\sum_{j>i} d_{ij|s'}v_{ij} + \left\{\sum_{i\in s'}\sum_{j>i} u_{ij} - \sum_{i\in s}\sum_{j>i} d_{ij|s'}u_{ij}\right\}\hat{B}_S\right]$$

where $\hat{B}_S = \left\{\sum_{i\in s}\sum_{j>i} d_{ij|s'}q_{ij}u_{ij}v_{ij}\right\}/\left\{\sum_{i\in s}\sum_{j>i} d_{ij|s'}q_{ij}u_{ij}^2\right\}$. For a general second-order finite population quantity such as Q, one needs to use $v_{ij} = \phi(y_i, y_j)$, $u_{ij} = E_\xi[\phi(y_i, y_j) \mid \boldsymbol{x}_i, \boldsymbol{x}_j]$, and replace $1/[n'(n'-1)]$ by $N(N-1)/[n'(n'-1)]$ in the above formulation.

If simple random sampling without replacement (SRSWOR) is also used at the second phase, $\hat{S}_{MC}^2$ reduces to

$$\hat{S}_{MC}^2 = s_Y^2 + \left[\hat{\boldsymbol{\theta}}^T(\boldsymbol{s}_X'^2 - \boldsymbol{s}_X^2)\hat{\boldsymbol{\theta}} + \hat{\sigma}^2\left(\frac{1}{n'}\sum_{i\in s'} v^2(\boldsymbol{x}_i) - \frac{1}{n}\sum_{i\in s} v^2(\boldsymbol{x}_i)\right)\right]\hat{B}_S \tag{10}$$

where s_Y^2 is the usual sample variance based on y_i, $i \in s$, and $\boldsymbol{s}_X'^2$ and $\boldsymbol{s}_X^2$ are the sample variance-covariance matrices for the $\boldsymbol{x}$ variable based on s' and s, respectively. If further the regression model (9) has a homogeneous variance structure, i.e., $v(\boldsymbol{x}_i) \equiv 1$, we have $\hat{S}_{MC}^2 = s_Y^2 + \hat{\boldsymbol{\theta}}^T(\boldsymbol{s}_X'^2 - \boldsymbol{s}_X^2)\hat{\boldsymbol{\theta}}\hat{B}_S$.

3. More Efficient Variance Estimators for the Regression Estimator

Under the commonly used regression model (9), it can be shown that $\hat{B}_Y = 1$ and the optimal model-calibration estimator for the population mean $\bar{Y} = N^{-1}\sum_{i=1}^N y_i$ is identical to the conventional regression estimator under two-phase sampling (Wu and Sitter 2001):

$$\hat{\bar{Y}}_{MC} = \frac{1}{n'}\left[\sum_{i\in s} d_{i|s'}y_i + \left(\sum_{i\in s'}\boldsymbol{x}_i - \sum_{i\in s} d_{i|s'}\boldsymbol{x}_i\right)^T\hat{\boldsymbol{\theta}}\right]$$

where $\hat{\boldsymbol{\theta}}$ are the estimated regression coefficients. If SRSWOR is used at the second phase, the approximate design-based variance of $\hat{\bar{Y}}_{MC}$ is given by

$$V_p(\hat{\bar{Y}}_{MC}) \doteq \left(\frac{1}{n'} - \frac{1}{N}\right)S_Y^2 + \left(\frac{1}{n} - \frac{1}{n'}\right)S_E^2$$

where S_Y^2 is defined as before, S_E^2 is the finite population variance defined over the residual variable $e_i = y_i - \boldsymbol{x}_i^T\boldsymbol{\theta}_N$, and $\boldsymbol{\theta}_N$ is the finite population regression coefficients. The conventional variance estimator for $\hat{\bar{Y}}_{MC}$ is obtained if one replaces S_Y^2 by s_Y^2 and S_E^2 by s_E^2, the sample variances based on the second-phase sample s, and using $\hat{e}_i = y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\theta}}$. See for example Cochran (1977, p. 343). We denote this basic variance estimator by $v_0(\hat{\bar{Y}}_{MC})$.

An improved variance estimator which makes more complete use of the large first-phase sample measured over the covariates $\boldsymbol{x}$ can be easily developed as follows. The S_Y^2 can be more efficiently estimated by $\hat{S}_{MC}^2$ given by (10). As for S_E^2, the optimal model-calibration estimator can be similarly constructed by noting that $E_\xi(e_i) = 0$ and

$E_\xi[(e_i - e_j)^2 \mid \boldsymbol{x}_i, \boldsymbol{x}_j] \doteq \sigma^2[v^2(\boldsymbol{x}_i) + v^2(\boldsymbol{x}_j)]$. The resulting estimator for S_E^2 is given by

$$\hat{S}_E^2 = s_E^2 + \left[\frac{1}{n'}\sum_{i\in s'} v^2(\boldsymbol{x}_i) - \frac{1}{n}\sum_{i\in s} v^2(\boldsymbol{x}_i)\right]\hat{B}_E$$

where $\hat{B}_E$ is similarly defined as $\hat{B}_S$ but using $v_{ij} = (\hat{e}_i - \hat{e}_j)^2$ and $u_{ij} = v^2(\boldsymbol{x}_i) + v^2(\boldsymbol{x}_j)$. It is interesting to notice that, if $v(\boldsymbol{x}_i) \equiv 1$, $\hat{S}_E^2 = s_E^2$, the large first-phase sample on $\boldsymbol{x}$ contains no extra information to improve the estimation of S_E^2. See Wu (2002) for more discussion on this issue.

Our proposed variance estimator is as follows:

$$\begin{aligned} v_1(\hat{\bar{Y}}_{MC}) = {} & \left(\frac{1}{n'} - \frac{1}{N}\right)s_Y^2 + \left(\frac{1}{n} - \frac{1}{n'}\right)s_E^2 + \left(\frac{1}{n'} - \frac{1}{N}\right)\hat{\boldsymbol{\theta}}^T(\boldsymbol{s}_X'^2 - \boldsymbol{s}_X^2)\hat{\boldsymbol{\theta}}\hat{B}_S \\ & + \left[\frac{1}{n'}\sum_{i\in s'} v^2(\boldsymbol{x}_i) - \frac{1}{n}\sum_{i\in s} v^2(\boldsymbol{x}_i)\right]\left[\left(\frac{1}{n'} - \frac{1}{N}\right)\hat{B}_S + \left(\frac{1}{n} - \frac{1}{n'}\right)\hat{B}_E\right]\hat{\sigma}^2 \end{aligned}$$

If $v(\boldsymbol{x}_i) \equiv 1$, the very last term in $v_1(\hat{\bar{Y}}_{MC})$ vanishes, and in this case $v_1(\hat{\bar{Y}}_{MC})$ reduces to $v_1(\bar{y}_{lr})$ proposed by Sitter (1997) if we also replace $\hat{B}_S$ by 1. Note that, under the model (9), $E_\xi(\hat{B}_S) \doteq 1$, $E_\xi(\hat{B}_E) \doteq 1$, an alternative variance estimator is obtained if we replace both $\hat{B}_S$ and $\hat{B}_E$ by 1. This is given by

$$\begin{aligned} v_2(\hat{\bar{Y}}_{MC}) = {} & \left(\frac{1}{n'} - \frac{1}{N}\right)s_Y^2 + \left(\frac{1}{n} - \frac{1}{n'}\right)s_E^2 + \left(\frac{1}{n'} - \frac{1}{N}\right)\hat{\boldsymbol{\theta}}^T(\boldsymbol{s}_X'^2 - \boldsymbol{s}_X^2)\hat{\boldsymbol{\theta}} \\ & + \left(\frac{1}{n} - \frac{1}{N}\right)\left[\frac{1}{n'}\sum_{i\in s'} v^2(\boldsymbol{x}_i) - \frac{1}{n}\sum_{i\in s} v^2(\boldsymbol{x}_i)\right]\hat{\sigma}^2 \end{aligned}$$

Both v_1 and v_2 are design-consistent. The design-based finite sample performances of v_1 and v_2 are further investigated in Section 5 through a simulation study. Also included in the simulation study is the delete-1 jackknife variance estimator, v_J. The detailed formulation of v_J was given by Sitter (1997).

4. Variance Estimation for the Distribution Function

4.1. Analytical variance estimators

Analytical variance estimators for $\hat{F}_{ME}(t)$ which make more efficient use of the first-phase sample can be developed in the same spirit to v_1 or v_2. Following the arguments of Chen and Wu (2002), it can be shown that

$$\begin{aligned} \hat{F}_{ME}(t) = {} & \frac{1}{\hat{n}'}\sum_{i\in s} d_{i|s'} I(y_i \le t) \\ & + \left[\frac{1}{n'}\sum_{i\in s'} g(\boldsymbol{x}_i, t) - \frac{1}{\hat{n}'}\sum_{i\in s} d_{i|s'} g(\boldsymbol{x}_i, t)\right]B_N + o_p\left(\frac{1}{\sqrt{n}}\right) \end{aligned} \tag{11}$$

where $\hat{n}' = \sum_{i\in s} d_{i|s'}$, $B_N = \sum_{i=1}^N [g(\boldsymbol{x}_i, t) - \bar{g}_N] I(y_i \le t) / \sum_{i=1}^N [g(\boldsymbol{x}_i, t) - \bar{g}_N]^2$, and $\bar{g}_N = N^{-1}\sum_{i=1}^N g(\boldsymbol{x}_i, t)$. Under two-phase sampling with SRSWOR at both phases,

we have

$$V_p[\hat{F}_{ME}(t)] \doteq \left(\frac{1}{n'} - \frac{1}{N}\right) S_I^2 + \left(\frac{1}{n} - \frac{1}{n'}\right) S_D^2$$

where S_I^2 and S_D^2 are the finite population variances defined over the variable $I_i = I(y_i \leq t)$ and $D_i = I(y_i \leq t) - g(\boldsymbol{x}_i, t)B_N$, respectively.

The usual substitution estimator for $V_p[\hat{F}_{ME}(t)]$ is denoted by $v_0[\hat{F}_{ME}(t)]$, with S_I^2 replaced by s_I^2 and S_D^2 replaced by s_D^2, the sample variances based on s.

A more efficient variance estimator is readily available if we estimate both S_I^2 and S_D^2 using the general method described in Section 2.3. The resulting estimators $\hat{S}_I^2$ and $\hat{S}_D^2$ can both be expressed as

$$\hat{S}^2 = s^2 + \left[\frac{1}{n'(n'-1)} \sum_{i \in s'} \sum_{j>i} u_{ij} - \frac{1}{n(n-1)} \sum_{i \in s} \sum_{j>i} u_{ij}\right] \hat{B}_F$$

where $\hat{B}_F = \sum_{i \in s} \sum_{j>i} u_{ij} v_{ij} / \sum_{i \in s} \sum_{j>i} u_{ij}^2$ and s^2 is the corresponding sample variance based on s. For S_I^2, $v_{ij} = (I_i - I_j)^2$, $u_{ij} = E_\xi[(I_i - I_j)^2 \mid \boldsymbol{x}_i, \boldsymbol{x}_j] = g_i + g_j - 2g_i g_j$, and $g_i = g(\boldsymbol{x}_i, t)$; for S_D^2, $v_{ij} = (D_i - D_j)^2$, $u_{ij} = E_\xi[(D_i - D_j)^2 \mid \boldsymbol{x}_i, \boldsymbol{x}_j] \doteq g_i(1 - g_i) + g_j(1 - g_j)$. As usual, any unknown model parameters appearing in D_i or u_{ij} will be replaced by appropriate sample-based estimates. Let $v_1[\hat{F}_{ME}(t)]$ be the estimator of $V_p[\hat{F}_{ME}(t)]$ obtained by using $\hat{S}_I^2$ and $\hat{S}_D^2$ in place of S_I^2 and S_D^2, respectively.

Note that $S_I^2 = N(N-1)^{-1} F_Y(t)[1 - F_Y(t)]$, an alternative estimator for S_I^2 could simply be $N(N-1)^{-1}\hat{F}_{ME}(t)[1 - \hat{F}_{ME}(t)]$. We denote the resulting estimator of $V_p[\hat{F}_{ME}(t)]$ as $v_2[\hat{F}_{ME}(t)]$, where S_D^2 is estimated by $\hat{S}_D^2$, as in $v_1[\hat{F}_{ME}(t)]$. Similar to the mean case, both v_1 and v_2 are design-consistent estimators of $V_p[\hat{F}_{ME}(t)]$.

4.2. *Jackknife variance estimator*

The pseudo-empirical maximum likelihood estimator $\hat{F}_{ME}(t)$ is a nonlinear estimator and its exact design-based variance does not have a closed form. The conventional delete-1 jackknife variance estimator often provides an attractive alternative in such cases.

Under two-phase sampling, the delete-1 estimator $\hat{F}_{ME}[j]$ needs to be computed differently for $j \in s$ and for $j \in s' - s$. Let $s(j)$ denote the set s with the jth unit deleted. Let $\hat{F}_{ME}[j] = \sum_{i \in s(j)} \hat{p}_i I(y_i \leq t)$ if $j \in s$, where the weights $\hat{p}_i$ $(i \neq j)$ maximize $\hat{l}(\boldsymbol{p}) = \sum_{i \in s(j)} d_{i|s'} \log(p_i)$ subject to $\sum_{i \in s(j)} p_i = 1$ $(0 < p_i < 1)$ and $\sum_{i \in s(j)} p_i g(\boldsymbol{x}_i, t) = (n'-1)^{-1} \sum_{i \in s'(j)} g(\boldsymbol{x}_i, t)$; let $\hat{F}_{ME}[j] = \sum_{i \in s} \hat{p}_i I(y_i \leq t)$ if $j \in s' - s$, where the weights $\hat{p}_i$ maximize $\hat{l}(\boldsymbol{p}) = \sum_{i \in s} d_{i|s'} \log(p_i)$ subject to $\sum_{i \in s} p_i = 1$ $(0 < p_i < 1)$ and $\sum_{i \in s} p_i g(\boldsymbol{x}_i, t) = (n'-1)^{-1} \sum_{i \in s'(j)} g(\boldsymbol{x}_i, t)$. The usual jackknife variance estimator is

$$v_J[\hat{F}_{ME}(t)] = \frac{n'-1}{n'} \sum_{j \in s'} \{\hat{F}_{ME}[j] - \hat{F}_{ME}(t)\}^2 \tag{12}$$

This variance estimator is consistent if SRSWOR is used at both phases and the first-phase sampling fraction n'/N is negligible, since $\hat{F}_{ME}(t)$ is asymptotically equivalent to a regression type estimator given by (11) (Sitter 1997). If the first-phase sampling fraction cannot be ignored, an ad hoc adjustment can be made by multiplying v_J by $1 - n'/N$. This has been done in the simulation study.

The major advantage of using v_J is the operational convenience. For parameters in a general form of $W = W\{F_Y(t)\}$, the jackknife variance estimator for $\hat{W} = W\{\hat{F}_{ME}(t)\}$ can be readily formulated by replacing $\hat{F}_{ME}[j]$ and $\hat{F}_{ME}(t)$ in (12) by $\hat{W}[j] = W\{\hat{F}_{ME}[j]\}$ and $\hat{W} = W\{\hat{F}_{ME}(t)\}$. This is most useful when the method is applied to infinite populations where parameters of interest are often in the form of $\theta = \theta(F)$.

5. Some Empirical Results

In this section we report some simulation results regarding the design-based finite sample performances of the proposed estimators using real survey data from the 1996 Statistics Canada Family Expenditure (FAMEX) Survey for the province of Ontario. The data set contains 2,396 observations over a variety of variables. In the simulation, the variable y, total expenditure, was used as the response, and x_1 (number of people in the household) and x_2 (total income after taxes) were included as auxiliary variables. It was found that a simple linear regression model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ gives a reasonable fit to the data, with the nonhomogeneous variance structure $V_\xi(\varepsilon) = x_2\sigma^2$ fitting slightly better than the constant variance model.

Treating the data set as a fixed finite population, we first selected a two-phase sample with SRSWOR at both phases, and then computed various estimators presented in Sections 2, 3, and 4. For the distribution function $F_Y(t)$ we included results on five different values of t corresponding to the 10th, 25th, 50th, 75th, and 90th population quantiles. The process was repeated $B = 5{,}000$ times for point estimators $\hat{\bar{Y}}_{MC}$, $\hat{S}^2_{MC}$ and $\hat{F}_{ME}(t)$, and $B = 1{,}000$ times for variance estimators v_0, v_1, v_2, and v_J.

The performance of an estimator $\hat{T}$ for a certain population quantity T was evaluated using the relative percentage bias ($RB\%$) and the relative efficiency (RE) defined as

$$RB\% = \frac{1}{B}\sum_{b=1}^{B}\frac{\hat{T}_b - T}{T}\times 100 \quad \text{and} \quad RE = \frac{MSE(\hat{T}_0)}{MSE(\hat{T})}$$

where $\hat{T}_b$ was computed from the bth simulated sample, $MSE(\hat{T}) = B^{-1}\sum_{b=1}^{B}(\hat{T}_b - T)^2$, and $\hat{T}_0$ denotes the baseline estimator for comparison. For the estimation of $\bar{Y}$, S_Y^2 and $F_Y(t)$, $\hat{T}_0$ is given by $\bar{y} = n^{-1}\sum_{i\in s} y_i$, $s_Y^2 = (n-1)^{-1}\sum_{i\in s}(y_i - \bar{y})^2$ and $\hat{F}_Y(t) = n^{-1}\sum_{i\in s} I(y_i \le t)$, respectively.

Table 1 reports the simulated RE's for the optimal-calibration or the model-calibrated pseudo-empirical maximum likelihood estimators for the population mean $\bar{Y}$, the population variance S_Y^2 and the finite population distribution function $F_Y(t)$ at $t = t_\alpha$, $\alpha = 0.10$, 0.25, 0.50, 0.75, and 0.90. The first-phase sample size was chosen as $n' = 400$. The absolute values of the RB's are all less than 2% and are not reported here to save space.

Table 1. Relative efficiency of estimators for $\bar{Y}$, S_Y^2 and $F_Y(t)$

n'	n	$\hat{\bar{Y}}_{MC}$	$\hat{S}^2_{MC}$	$\hat{F}_{ME}(t)$ at $t = t_\alpha$				
				0.10	0.25	0.50	0.75	0.90
400	40	3.30	1.32	1.61	2.30	2.34	1.98	1.45
	80	2.73	1.33	1.70	2.05	2.13	1.87	1.61
	160	2.01	1.22	1.42	1.69	1.72	1.59	1.45

Table 2. Relative bias (in %) of variance estimators for $\hat{\bar{Y}}_{MC}$ and $\hat{F}_{ME}(t)$

n'	n	v	$\hat{\bar{Y}}_{MC}$	$\hat{F}_{ME}(t)$ at $t = t_\alpha$				
				0.10	0.25	0.50	0.75	0.90
400	40	v_0	−1.7	−23.6	−4.1	−4.9	−3.5	−22.4
		v_1	−2.1	−24.5	−3.2	−4.5	−3.7	−23.4
		v_2	−2.1	−24.6	−3.4	−4.7	−4.0	−23.6
		v_J	1.7	16.4	3.8	−7.1	−3.6	61.8
	80	v_0	−4.5	−9.7	−2.8	−3.6	−0.2	−6.6
		v_1	−4.1	−10.5	−2.9	−3.3	−0.1	−6.8
		v_2	−4.3	−10.7	−3.1	−3.5	−0.3	−7.0
		v_J	−9.2	1.5	−7.0	−9.1	−6.8	−0.9
	160	v_0	0.5	−4.8	0.3	−0.6	−0.4	−3.2
		v_1	0.8	−4.3	0.5	−0.7	−0.4	−3.1
		v_2	0.5	−4.5	0.3	−0.9	−0.6	−3.4
		v_J	−2.3	−5.6	−4.3	−5.6	−5.9	−7.4

The proposed estimators perform uniformly better and are much superior to the conventional estimator $\hat{T}_0$ in all cases. With fixed first-phase sample size, the gain in efficiency seems larger when the second-phase sample size n is smaller, although such a trend is not monotonic in terms of sample size. As pointed out by one of the referees, this is due to the fact that $V_p(\hat{T}) \to V_p(\hat{T}_0)$ as $n \to n'$ for fixed n'. As for the distribution function $F_Y(t)$, the improvement is more dramatic when t is not in the tail regions.

For variance estimators, the true values (i.e., T in the definitions of *RB* and *RE*) of $V_p(\hat{\bar{Y}}_{MC})$ and $V_p(\hat{F}_{ME}(t))$ were estimated from another 5,000 independent simulation runs. The simulated *RB*'s of v_0, v_1, v_2 and v_J are reported in Table 2. All *RB*'s are less than 11%, except for those cases relating to the estimation of the variance of $\hat{F}_{ME}(0.10)$ and $\hat{F}_{ME}(0.90)$ when $n = 40$. The *RB*'s in these cases all take negative values and are smaller than −20% for v_0, v_1 and v_2. The jackknife estimator v_J also significantly overestimates the variance in these cases. It seems that none of these estimators, including the naive v_0, provides acceptable estimates when n is small and t is in the tail region.

The simulated *RE*'s for the variance estimators of $\hat{\bar{Y}}_{MC}$ and $\hat{F}_{ME}(t)$ are presented in Table 3. Note that we excluded the *RE*'s for the case of $n = 40$ at $t = t_{0.10}$ and $t_{0.90}$ where

Table 3. Relative efficiency of variance estimators for $\hat{\bar{Y}}_{MC}$ and $\hat{F}_{ME}(t)$

n'	n	v	$\hat{\bar{Y}}_{MC}$	$\hat{F}_{ME}(t)$ at $t = t_\alpha$				
				0.10	0.25	0.50	0.75	0.90
400	40	v_1	1.16	—	1.18	1.08	1.29	—
		v_2	1.12	—	1.19	1.09	1.30	—
		v_J	0.85	—	0.22	0.87	0.28	—
	80	v_1	1.14	1.19	1.20	1.06	1.42	1.40
		v_2	1.14	1.21	1.21	1.06	1.42	1.42
		v_J	1.26	0.12	0.76	0.66	0.85	0.11
	160	v_1	1.14	1.31	1.23	1.07	1.44	1.37
		v_2	1.16	1.34	1.25	1.09	1.45	1.38
		v_J	1.12	0.61	0.86	0.58	0.96	1.05

the bias is not negligible, since comparison in terms of *RE* under such cases is probably misleading. The variance estimator for baseline comparison is v_0.

Table 3 can be summarized as follows: (i) v_1 and v_2 perform similar to each other and are uniformly better than v_0; (ii) the improvement from using v_1 and v_2 over v_0 is only marginal for estimating $V_p(\hat{\bar{Y}}_{MC})$; (iii) v_1 and v_2 seem more efficient for estimating $V_p(\hat{F}_{ME}(t))$ for t at large quantiles; and (iv) the jackknife variance estimator, which does not reuse the auxiliary information available at the first phase, is less stable and less efficient in most cases. One explanation for (ii) is probably the fact that the linear model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ with the nonhomogeneous variance structure $V_\xi(\varepsilon) = x_2 \sigma^2$ only provides a slightly better fit than the constant variance model. Under such circumstances there is little room to improve the estimation of S_E^2 by using auxiliary information through a nonhomogeneous variance structure. The marginal gain comes from the estimation of S_Y^2 using $\hat{S}_{MC}^2$. When the variance structure is clearly nonhomogeneous, higher efficiency can be expected from using v_1 and v_2. See the theoretical and empirical results reported in Wu (2002) for the case of known complete auxiliary information.

6. Applications to Measurement Error and Nonresponse

The proposed optimal calibration estimators under two-phase sampling are applicable to estimation problems under measurement errors or nonresponse. Let y_i be the true value of a characteristic of interest for the ith unit in the finite population. Suppose it is difficult or very expensive to obtain exact measurement but an inaccurate value (i.e., measurement with error), say z_i, of y for the ith unit, can be obtained quite easily. In such situations the two-phase sampling technique is often employed. A large first-phase sample s' is taken and the cheap inaccurate measurements z_i are collected for all $i \in s'$. A smaller second-phase sample $s \subset s'$ is drawn and the exact measurements y_i are collected using more extensive effort. The goal is to estimate various finite population quantities defined over y_i.

The relationship between the y_i's and the z_i's can be described by the so-called regression calibration model (Carroll et al. 1995),

$$y_i = \alpha + \beta z_i + \varepsilon_i, \quad i = 1, 2, \ldots, N \tag{13}$$

where the ε_i's are assumed independent and identically distributed random variates with $E_m(\varepsilon_i) = 0$ and $V_m(\varepsilon_i) = \sigma^2$, and the subscript m refers to the measurement error model (13). If we treat z_i as an auxiliary variable, the methodology developed in Sections 2, 3, and 4 can be applied directly here for estimation under the assumed measurement error model.

Item nonresponse is often handled through imputation. If the response variable y is highly correlated with certain covariate(s) $\boldsymbol{x}$ and missing values do not occur for the $\boldsymbol{x}$ variables, a direct analysis based on the observed sample data using our proposed estimation strategy may be more desirable. The original sample with complete information on $\boldsymbol{x}$ can be viewed as a first-phase sample, and the subsample with observed responses on y is treated as a second-phase sample. The loss of information due to the nonresponse can be retreated from auxiliary information through the optimal estimation strategy. The

inference is conditionally valid for the given sample. An advantage of this approach is that the resulting optimal calibration estimators are design-consistent even if the super-population model (2) is misspecified. This is in contrast to estimation based on imputed data where the validity of the results depends on the correctness of the model used for imputation.

7. Concluding Remarks

The calibration method has gained much popularity since its emergence in the early 90s, and calibration estimators are routinely computed by many survey organizations. The optimal calibration approach provides a unified framework for the efficient estimation of various finite population quantities when complete auxiliary information is available. It has been demonstrated in this article that the method also provides a flexible and efficient way of using auxiliary information at the estimation stage under a two-phase sampling scheme. The proposed optimal calibration estimators for a second-order finite population quantity such as the population variance can perfectly be used to obtain more efficient variance estimators for a first-order finite population quantity such as the total or the distribution function. The gain in efficiency can be appreciable even for a first-phase sample with moderate size. Applications to measurement error problems or nonresponse are straightforward. The method has the potential to be applied to estimation problems under infinite populations where a two-phase sample is available.

8. References

Carroll, R.J., Ruppert, D., and Stefanski, L.A. (1995). Measurement Errors in Non-Linear Models. Chapman & Hall.

Chen, J. and Wu, C. (2002). Estimation of Distribution Function and Quantiles Using the Model-calibrated Pseudo Empirical Likelihood Method. Statistica Sinica, 12, 1223–1239.

Chen, J., Sitter, R.R., and Wu, C. (2002). Using Empirical Likelihood Method to Obtain Range Restricted Weights in Regression Estimators for Surveys. Biometrika, 89, 230–237.

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

Deville, J.C. and Särndal, C.E. (1992). Calibration Estimators in Survey Sampling. Journal of the American Statistical Association, 87, 376–382.

Luan, Y. (2001). Model-Calibration Approach under Two-Phase Sampling. Unpublished master's essay, Department of Statistics and Actuarial Science, University of Waterloo, Canada.

Sitter, R.R. (1997). Variance Estimation for the Regression Estimator in Two-Phase Sampling. Journal of the American Statistical Association, 92, 780–787.

Särndal, C.E., Swensson, B., and Wretman, J. (1992). Model Assisted Survey Sampling. Springer, New York.

Wu, C. (2002). Optimal Calibration Estimators in Survey Sampling. Working paper 2002-01. Department of Statistics and Actuarial Science, University of Waterloo, Canada.

Wu, C. and Sitter, R.R. (2001). A Model-calibration Approach to Using Complete Auxiliary Information from Survey Data. Journal of the American Statistical Association, 96, 185–193.

Received March 2002
Revised November 2002