Biometrika (2003), **90,** 4, *pp.* 937–951

Printed in Great Britain

Optimal calibration estimators in survey sampling

By CHANGBAO WU

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

cbwu@uwaterloo.ca

Summary

We show that the model-calibration estimator for the finite population mean, which was proposed by Wu & Sitter (2001) through an intuitive argument, is optimal among a class of calibration estimators. We also present optimal calibration estimators for the finite population distribution function, the population variance, the variance of a linear estimator and other quadratic finite population functions under a unified framework. The proposed calibration estimators are optimal under the true model but remain design consistent even if the working model is misspecified. A limited simulation study shows that the improvement of these optimal estimators over the conventional ones can be substantial. The question of when and how auxiliary information can be used for both the estimation of the population mean using a generalised regression estimator and the estimation of its variance through calibration is addressed clearly under the proposed general methodology. Some fundamental issues in using auxiliary information from survey data are also addressed in the context of optimal estimation.

Some key words: Asymptotic design variance; Auxiliary information; Model calibration; Optimal estimation; Superpopulation.

1. Introduction

The notion of calibration estimators was introduced by Deville & Särndal (1992) in the context of using auxiliary information from survey data. Suppose $U = \{1, 2, \ldots, N\}$ is the set of labels for the finite population. Let (y_i, x_i) be the values of the study variable y and the vector of auxiliary variables x attached to the ith unit. The question is how to estimate $\bar{Y} = N^{-1} \sum_{i=1}^{N} y_i$ effectively using the known population totals $X = \sum_{i=1}^{N} x_i$ at the estimation stage. Let $s = \{1, 2, \ldots, n\}$ be the set of sampled units under a general sampling design, p, and let $\pi_i = \mathrm{pr}(i \in s)$ be the first-order inclusion probabilities. The conventional calibration estimator for $\bar{Y}$ is defined by $\tilde{Y}_C = N^{-1} \sum_{i \in s} w_i y_i$, where the w_i's are modified from the basic design weights $d_i = 1/\pi_i$ by minimising a distance measure Φ_s between the w_i's and the d_i's subject to constraints

$$\sum_{i \in s} w_i x_i = \sum_{i=1}^{N} x_i. \tag{1·1}$$

The most commonly used distance measure is the chi-squared distance

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

where the q_i's are known positive constants uncorrelated with the d_i's. Alternative distance measures can also be considered; see Deville & Särndal (1992) for a detailed discussion.

There are two basic components in the construction of calibration estimators, namely a distance measure and a set of calibration equations. The choice of a distance measure is less critical in terms of efficiency since the resulting estimators are all asymptotically equivalent to the one obtained by using a chi-squared distance with a certain choice of q_i's (Deville & Särndal, 1992). Calibration equations (1·1) are routinely used by many survey organisations and are referred to as benchmark constraints. Benchmark constraints are often imposed in practice for two reasons: the surveyor may believe that the weights which give perfect estimates for the auxiliary variables should also give a good estimate for the study variable; and the auxiliary information may only be available at the aggregate level, i.e. only X is known. Statistics practitioners in areas such as demography sometimes insist on benchmarking over lots of variables to match the known totals from a census at the risk of worsening the efficiency of the estimators. On the other hand, if complete auxiliary information $x_1, \ldots, x_N$ is known, which is often the case in many survey problems, a very compelling question to ask would be 'what is the best calibration equation to be used in the construction of a calibration estimator?'

Let $u_i = u(x_i)$ $(i = 1, \ldots, N)$, where $u(.)$ is a real-valued function. If we replace (1·1) by

$$\sum_{i \in s} w_i u(x_i) = \sum_{i=1}^{N} u(x_i), \tag{1·2}$$

then the question becomes 'which $u(.)$ will make $\tilde{Y}_C$ most efficient?' Note that the benchmark constraints (1·1) consist of k equations, where k is the number of components in x, while constraint (1·2) only has one equation involving the single data-reduction variable $u = u(x)$. The single calibration equation (1·2) is indeed more general than the k constraints (1·1), because of the unspecified function $u(.)$. For any k-dimensional vector $x = (x_1, \ldots, x_k)$, if we use $u(x) = \theta_0 + \theta_1 x_1 + \ldots + \theta_k x_k$, where $\theta = (\theta_0, \ldots, \theta_k)$ are estimated by ordinary least squares, then the calibration estimator of the population mean or total obtained by using the single constraint (1·2) is identical to the one using (1·1) (Wu & Sitter, 2001, Theorem 1). The conventional calibration estimator based on (1·1) is therefore a special member of the class of estimators considered in this paper.

It is well known that in survey sampling a uniformly minimum variance unbiased estimator does not exist under the design-based framework. Indeed the only choice of $u(.)$ that results in a $\tilde{Y}_C$ with minimum variance is $u(x_i) \equiv y$, and this of course is practically useless.

The model-assisted optimal estimators that minimise the expected design variance $E_\xi\{V_p(\tilde{Y})\}$ under a superpopulation model have been discussed by several authors; see for example the work by Godambe (1955), Godambe & Thompson (1973), Cassel et al. (1976) and Isaki & Fuller (1982). The expected design variance was also termed 'anticipated variance' by Isaki & Fuller (1982). Note that E_p and V_p refer to the expectation and variance under the sampling design, p, and E_ξ and V_ξ denote the expectation and variance under a superpopulation model, ξ.

In this paper, we use a similar criterion. Calibration estimators belong to the class of nonlinear estimators and their exact design variance or mean squared error does not have a closed form. A natural replacement for optimality considerations is to minimise the model expectation of the asymptotic design variance $E_\xi\{AV_p(\tilde{Y})\}$, where AV_p represents the design-based asymptotic variance. Since the bias $B_p(\tilde{Y}_C) = E_p(\tilde{Y}_C - \bar{Y})$ of a calibra-

tion estimator $\tilde{Y}_C$ satisfies $B_p(\tilde{Y}_C) = o(n^{-\frac{1}{2}})$ and $V_p(\tilde{Y}_C) = O(n^{-1})$, minimising $E_\xi\{AV_p(\tilde{Y}_C)\}$ is equivalent to minimising $E_\xi\{E_p(\tilde{Y}_C - \bar{Y})^2\}$ asymptotically.

In § 2, we show that the model-calibration estimator for the finite population mean, which was proposed by Wu & Sitter (2001) through an intuitive argument, is indeed optimal among a class of calibration estimators in the sense of minimising the expected asymptotic design variance under a superpopulation model and any regular sampling design. The result provides a unified framework for constructing optimal calibration estimators for the finite population distribution function, the population variance, the variance of a linear estimator and other quadratic finite population functions. Optimal calibration estimators for the distribution function are presented in § 3, and estimators for a general second-order finite population quantity using optimal calibration are constructed in § 4. Also in § 4, the question of when and how auxiliary information can be used for both the estimation of the population mean using a generalised regression estimator and the estimation of its variance through calibration is addressed clearly under the unified framework. The optimal pseudo empirical maximum likelihood estimators, which are asymptotically equivalent to the optimal calibration estimators, are particularly useful in estimating the distribution function, the population variance and other known nonnegative quantities. Our proposed estimators are optimal under the true model but remain design consistent even if the working model is misspecified. Results of a limited simulation study on the performance of these optimal estimators under the true model and the robustness of these estimators against model misspecifications along with comparison to the conventional estimators are reported in § 5. Some fundamental issues in using auxiliary information from survey data are also addressed under this framework, and these together with some concluding remarks are given in § 6.

2. The optimality of the model-calibration estimator

For asymptotic analysis, we assume there is a sequence of finite populations, indexed by v. The population size and sample size for the vth population are denoted by N_v and n_v. As $v \to \infty$, $N_v \to \infty$ and $n_v \to \infty$. All limiting processes should be understood to mean $v \to \infty$. The index v will be suppressed to simplify notation. For a detailed formulation of this asymptotic framework, see Isaki & Fuller (1982).

We consider situations where the finite population measurements $\{(x_i, y_i), i = 1, \ldots, N\}$ can be viewed as independent realisations from a superpopulation model ξ such that

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

where θ, typically vector-valued, and σ^2 are model parameters, and the mean function $\mu(.,.)$ and the variance function $v(.)$ have known forms. The $v(.)$ could also be a known function of $\mu_i = \mu(x_i, \theta)$ as in the case of a generalised linear model. We assume that complete auxiliary information $(x_1, \ldots, x_N)$ is available.

Let $\tilde{Y}_{C_u}$ be a calibration estimator of $\bar{Y}$ when $C_u = \{u(x_1), u(x_2), \ldots\}$ is used in (1·2) and an arbitrary distance measure is used. Let L be the set of sequences $C_u = \{u(x_1), u(x_2), \ldots\}$ for all conceivable functions $u(.)$ such that

$$N^{-1}\sum_{i=1}^{N}\{u(x_i)\}^6 = O(1)$$

and $N^{-1}\sum_{i=1}^{N}\{u(x_i)\}^2 \to c \neq 0$ as $N \to \infty$. These finite moment conditions on the sequence

$C_u \in L$ are not very restrictive and are used to furnish the proofs. We assume that $\{\mu(x_1, \theta), \mu(x_2, \theta), \ldots\} \in L$ and $\{v(x_1), v(x_2), \ldots\} \in L$.

A sampling design is said to be regular if the design results in a fixed sample size, has inclusion probabilities π_i and π_{ij} independent of the response measurements y_i given x_i, and satisfies the following conditions.

Condition 1. We require that $\max_{i \in s} nd_i/N = O(1)$.

Condition 2. We require that $N^{-1} \sum_{i \in s} d_i u_i - N^{-1} \sum_{i=1}^{N} u_i = O_p(n^{-\frac{1}{2}})$ for any sequence $(u_1, u_2, \ldots) \in L$.

Condition 1 simply states that no basic design weight is disproportionately large. Condition 2 can sufficiently be replaced by assuming that the Horvitz–Thompson estimator for $\bar{u}_N = N^{-1} \sum_{i=1}^{N} u_i$ is asymptotically normally distributed.

THEOREM 1. *Among the class of calibration estimators* $\tilde{Y}_{C_u}$ *with*

$$C_u = \{u(x_1), u(x_2), \ldots\} \in L,$$

the choice of $C_\mu = \{\mu(x_1, \theta), \mu(x_2, \theta), \ldots\}$ *minimises* $E_\xi\{AV_p(\tilde{Y}_{C_u})\}$ *under model* (2·1) *and any regular sampling design.*

See the Appendix for the proof.

In practice, the model parameter θ will have to be replaced by a sample-based estimator, $\hat{\theta}$. The resulting estimator $\tilde{Y}_{\mathrm{MC}}$ was termed by Wu & Sitter (2001) the model-calibration estimator of $\bar{Y}$. While $\tilde{Y}_{\mathrm{MC}}$ is optimal under the true model, it remains design-consistent even if the working model is misspecified. In other words, $\tilde{Y}_{\mathrm{MC}}$ is robust against model departures. To see this, from the proof of Theorem 1 in the Appendix, we have

$$\tilde{Y}_{\mathrm{MC}} = \frac{1}{N} \sum_{i \in s} d_i y_i + \frac{1}{N} \left\{ \sum_{i=1}^{N} \mu(x_i, \hat{\theta}) - \sum_{i \in s} d_i \mu(x_i, \hat{\theta}) \right\} \hat{B}, \tag{2·2}$$

where $\hat{B}$ is similarly defined as in the Appendix. Under the regularity conditions (i)–(iii) described in Wu & Sitter (2001, p. 187), it can be shown that

$$\tilde{Y}_{\mathrm{MC}} = \tilde{Y}_{\mathrm{HT}} + \frac{1}{N} \left\{ \sum_{i=1}^{N} \mu(x_i, \theta_N) - \sum_{i \in s} d_i \mu(x_i, \theta_N) \right\} B_N + o_p(n^{-\frac{1}{2}}), \tag{2·3}$$

where $\tilde{Y}_{\mathrm{HT}}$ is the Horvitz–Thompson estimator and θ_N and B_N are finite population parameters estimated by $\hat{\theta}$ and $\hat{B}$, respectively. Since $E_p\{\sum_{i=1}^{N} \mu(x_i, \theta_N) - \sum_{i \in s} d_i \mu(x_i, \theta_N)\} = 0$, the model-calibration estimator $\tilde{Y}_{\mathrm{MC}}$ will be design-consistent under any working model and sampling design satisfying the regularity conditions, with bias of the order of $o(n^{-\frac{1}{2}})$. When a nonlinear working model is used, however, care has to be given to the verification of these regularity conditions; see § 5 for an example based on the log-linear model.

The optimal calibration variable $\mu_i = \mu(x_i, \hat{\theta})$ depends on the response variable y through the estimated model parameters $\hat{\theta}$, and therefore the optimal calibration weights will also depend on the response variable. This dependence will restrict its applicability in some cases. It may be required to have a single set of weights not depending on the y-variables. There are also cases where the statistical agency publishes different weights for different purposes. The optimal weights computed for a particular y will also produce design-consistent estimators for other response variables, since the resulting estimator will still have the form of (2·2) but in this case the mean function $\mu(.,.)$ and the estimated parameters $\hat{\theta}$ will be associated with a different response variable. Design consistency of the estimator, however, can be easily argued through (2·3).

The optimal calibration approach can be applied to the pseudo empirical likelihood method of Chen & Sitter (1999). Let $\tilde{Y}_{\mathrm{EC}_u}$ be the pseudo empirical maximum likelihood estimator of $\bar{Y}$ obtained by calibrating over $C_u = \{u(x_1), u(x_2), \ldots\}$; that is, $\tilde{Y}_{\mathrm{EC}_u} = \sum_{i \in s} \hat{p}_i y_i$, where the $\hat{p}_i$'s maximise the pseudo empirical loglikelihood function $l(p) = \sum_{i \in s} d_i \log(p_i)$ subject to constraints

$$\sum_{i \in s} p_i = 1 \quad (0 < p_i < 1), \quad \sum_{i \in s} p_i u(x_i) = \frac{1}{N} \sum_{i=1}^{N} u(x_i). \tag{2·4}$$

The model-calibrated pseudo empirical maximum likelihood estimator $\tilde{Y}_{\mathrm{ME}}$ of $\bar{Y}$ is obtained when $C_\mu = \{\mu(x_1, \hat{\theta}), \mu(x_2, \hat{\theta}), \ldots\}$ is used in constraints (2·4).

THEOREM 2. *Among the class of pseudo empirical maximum likelihood estimators* $\tilde{Y}_{\mathrm{EC}_u}$ *for* $\bar{Y}$ *where* $C_u \in L$, *the choice of* $C_\mu = \{\mu(x_1, \theta), \mu(x_2, \theta), \ldots\}$ *minimises* $E_\xi\{AV_p(\tilde{Y}_{\mathrm{EC}_u})\}$ *under model* (2·1) *and any regular sampling design.*

See the Appendix for the proof.

The model-calibrated pseudo empirical maximum likelihood estimator for $\bar{Y}$ is asymptotically equivalent to the model-calibration estimator and is optimal within the same context. Simple algorithms for computing the estimator $\tilde{Y}_{\mathrm{ME}}$ have been developed by Chen et al. (2002). The most attractive feature of the estimator $\tilde{Y}_{\mathrm{ME}}$, however, is the intrinsic properties of the weights: $\hat{p}_i > 0$ and $\sum_{i \in s} \hat{p}_i = 1$. This is particularly useful when the method is extended to estimate the distribution function and other known nonnegative quantities. This is detailed in § 3 for the distribution function and in § 4 for the estimation of variance and other quadratic functions.

3. Optimal calibration estimators for the distribution function

The finite population distribution function $F_Y(t) = N^{-1} \sum_{i=1}^{N} I(y_i \leqslant t)$ is also a finite population mean defined for an indicator variable $z_i = I(y_i \leqslant t)$. Without using any auxiliary information, estimation of $F_Y(t)$ is a special case of estimating the population mean and is usually straightforward. In the presence of auxiliary information, the following aspects require attention.

(a) While benchmark constraints (1·1) calibrated directly over the x-variables are sometimes justifiable for the estimation of $\bar{Y}$, this internal consistency requirement, that the weights for the study variable provide perfect estimates for the auxiliary variables, is not needed for the estimation of $F_Y(t)$. Efficiency will be the primary concern.

(b) We have to work with the indicator variable $z_i = I(y_i \geqslant t)$ and to consider the issue of local efficiency, at a particular value of t, versus global efficiency, at an arbitrary t, in estimating $F_Y(t)$.

(c) It is desirable that an estimator of $F_Y(t)$, $\hat{F}_Y(t)$ say, be itself a distribution function, so that quantile estimates can be obtained through direct inversion of $\hat{F}_Y(t)$.

Many techniques for estimating $\bar{Y}$, when applied directly to the estimation of $F_Y(t)$, will produce unsatisfactory results. For instance, in the case of a scalar x variable, a regression-type estimator for $F_Y(t)$ will have the form $\hat{F}_{\mathrm{RE}}(t) = \hat{F}_Y(t) + \{F_X(t) - \hat{F}_X(t)\}\hat{B}$, where $\hat{F}_Y(t)$ and $\hat{F}_X(t)$ are Horvitz–Thompson-type estimators for $F_Y(t)$ and $F_X(t) = N^{-1} \sum_{i=1}^{N} I(x_i \leqslant t)$. The $\hat{B}$ is the estimated slope of regressing $I(y_i \leqslant t)$ on $I(x_i \leqslant t)$. The estimator $\hat{F}_{\mathrm{RE}}(t)$ suffers from several drawbacks, the obvious one being that $\hat{F}_{\mathrm{RE}}(t)$ is not a distribution function and can take values outside $[0, 1]$.

The model-calibrated pseudo empirical likelihood method provides estimators of

$F_Y(t)$ which are not only efficient but are also themselves genuine distribution functions. The optimal calibration variable $\mu(x_i, \theta)$ should now be replaced by $g(x_i, t) = E_\xi\{I(y_i \leqslant t)|x_i\} = \mathrm{pr}(y_i \leqslant t|x_i)$. Two types of working model can be considered for obtaining $g(x_i, t)$, models that relate the y_i to the x_i or models that relate the indicator variable $I(y_i \leqslant t)$ to the x_i. Under the commonly used regression model, we have

$$y_i = x_i'\theta + v(x_i)\varepsilon_i \quad (i = 1, 2, \ldots, N), \tag{3·1}$$

where the ε_i's are independent and identically distributed random variates with mean 0 and variance σ^2. Let $G(.)$ be the cumulative distribution function of the ε_i's. We have

$$g(x_i, t) = \mathrm{pr}(y_i \leqslant t|x_i) = G\{(t - x_i'\theta)/v(x_i)\}. \tag{3·2}$$

As in the mean case, the model parameter θ will have to be replaced by a sample-based design-consistent estimator in applications.

Note that $g_i = g(x_i, t)$ are probabilities. An alternative modelling process is to use a generalised linear model for the binary observations $I(y_i \leqslant t)$, such as a logistic regression model

$$\log\left(\frac{g_i}{1 - g_i}\right) = x_i'\theta, \tag{3·3}$$

with the usual variance function $V(g) = g(1 - g)$. Under model (3·3) we have $g(x_i, t) = \exp(x_i'\theta)/\{1 + \exp(x_i'\theta)\}$. Let $\hat{F}_{\mathrm{EC}_u}(t) = \sum_{i \in s}\hat{p}_i I(y_i \leqslant t)$, where the $\hat{p}_i$'s maximise $l(p)$ subject to constraints (2·4) with $C_u = \{u(x_1), u(x_2), \ldots\}$.

THEOREM 3. *The pseudo empirical maximum likelihood estimator $\hat{F}_{\mathrm{ME}}(t)$ calibrated over $\{g(x_1, t), g(x_2, t), \ldots\}$ is optimal among the class of estimators $\hat{F}_{\mathrm{EC}_u}(t)$ with $C_u \in L$ under the working model* (3·1) *or* (3·3) *and any regular sampling design.*

Proof. The result follows directly from Theorem 2 if one replaces y_i by $I(y_i \leqslant t)$ and $\mu(x_i, \theta)$ by $g(x_i, t)$. □

The design-based properties and small sample performance of these estimators and the related quantile estimation problem are investigated in Chen & Wu (2002). With complete auxiliary information, there exist several other estimators for the distribution function in the literature. Two leading competitors, the model-based estimator of Chambers & Dunstan (1986) and the model-assisted difference estimator of Rao et al. (1990), together with the calibration estimator of Chen & Sitter (1999) calibrated directly over the x variables, will be examined in the simulation study reported in § 5.

Note that the two working models (3·1) and (3·3) discussed above are not compatible with each other. Optimality of the resulting estimator, therefore, is meaningful under the chosen model. Note also that the optimal calibration variable $g(x_i, t)$ depends on t. No single set of weights $\hat{p}_i$ will produce an estimator that is optimal for all t. Chen & Wu (2002) suggest using a fixed t_0 in $g(x_i, t)$ while the resulting weights are used for any t in $\hat{F}_{\mathrm{ME}}(t)$. As a result $\hat{F}_{\mathrm{ME}}(t)$ is a genuine distribution function. Chen & Wu (2002) demonstrate through a simulation study that this $\hat{F}_{\mathrm{ME}}(t)$ is very efficient for values of t in a wide neighbourhood of t_0. The actual value of t_0 can be easily determined so as to maximise the efficiency of the resulting estimator when a certain neighbourhood of t_0 is of interest.

4. Optimal estimation of variance and other quadratic functions

4·1. *Estimation of second-order population quantities*

Estimation of variance and other second-order finite population quantities using auxiliary information has been addressed by many survey researchers. Various techniques have been attempted, such as regression, ratio and calibration estimation; see Sitter & Wu (2002) for a literature review. A common weakness of these approaches is the ad hoc argument of applying certain techniques, which were originally developed for estimating $\bar{Y}$, to estimate variance or other second-order population parameters without a common framework that unifies the two types of finite population parameter.

The optimal model-calibration and the model-calibrated pseudo empirical likelihood methods can be extended to handle variance and other second-order finite population parameters through a batch approach. For parameters in a general form of $Q = \sum_{i=1}^{N} \sum_{j=i+1}^{N} \phi(y_i, y_j)$, which include the population variance

$$S^2 = N^{-1} \sum_{i=1}^{N} (y_i - \bar{Y})^2 = \{N(N-1)\}^{-1} \sum_{i=1}^{N} \sum_{j=i+1}^{N} (y_i - y_j)^2$$

and the variance of the Horvitz–Thompson estimator

$$V_p(\hat{Y}_{\mathrm{HT}}) = \sum_{i=1}^{N} \sum_{j=i+1}^{N} (\pi_i \pi_j - \pi_{ij})(y_i/\pi_i - y_j/\pi_j)^2$$

as special cases, a unified estimation strategy can be developed as follows.

We may view Q as a total over a synthetic finite population, that is $Q = \sum_{\alpha=1}^{N^*} t_\alpha$, where $\alpha = (ij) = 1, 2, \ldots, N^*$, $t_\alpha = \phi(y_i, y_j)$ for $\alpha = (ij)$, and $N^* = N(N-1)/2$ is the total number of pairs. The sample data over the synthetic population consist of all pairs from the original sample: $s^* = \{(ij) : i < j, i, j \in s\}$. The 'first-order' inclusion probabilities under this setting are $\pi_{ij} = \mathrm{pr}(i, j \in s)$, and the 'basic design weights' are $d_{ij} = 1/\pi_{ij}$. The mean function $\mu(x_i, \theta) = E_\xi(y_i | x_i)$ should now be replaced by $E_\xi\{\phi(y_i, y_j) | x_i, x_j\}$.

If we use the original pair index (ij), the model-calibration estimator of Q is defined as $\hat{Q}_{\mathrm{MC}} = \sum_{i \in s} \sum_{j>i} w_{ij} \phi(y_i, y_j)$, where the weights w_{ij} minimise the modified chi-squared distance measure

$$\Phi_{s^*} = \sum_{i \in s} \sum_{j>i} (w_{ij} - d_{ij})^2/(d_{ij} q_{ij})$$

subject to

$$\sum_{i \in s} \sum_{j>i} w_{ij} E_\xi\{\phi(y_i, y_j) | x_i, y_j\} = \sum_{i=1}^{N} \sum_{j=i+1}^{N} E_\xi\{\phi(y_i, y_j) | x_i, x_j\}. \tag{4·1}$$

Let $\hat{Q}_{C_u^*}$ be a calibration estimator of Q when $C_u^* = \{u(x_i, x_j), i, j = 1, 2, \ldots\}$ is used in (4·1) as the calibration variable. Let L^* be the set of all possible sequences $C_u^* = \{u(x_i\ x_j), i, j = 1, 2, \ldots\}$ satisfying a finite moment condition similar to the one used in defining L. If we redefine the regular sampling design by replacing the d_i's in Conditions 1 and 2 of § 2 by d_{ij} with suitable reformulation, we have the following result.

Theorem 4. *Among the class of calibration estimators $\hat{Q}_{C_u^*}$ with*

$$C_u^* = \{u(x_i, x_j), i, j = 1, 2, \ldots\} \in L^*,$$

the model-calibration estimator $\hat{Q}_{\mathrm{MC}}$ attains the minimum value of $E_\xi\{AV_p(\hat{Q}_{C_u^*})\}$ under model (2·1) *and any regular sampling design.*

Proof. The result of Theorem 1 does not apply directly here because of a weak corre-

lation among the sequence of $t_\alpha = \phi(y_i, y_j)$, for $\alpha = (ij) = 1, 2, \ldots, N^*$, since t_α and $t_{\alpha'}$ are not independent of each other under model (2·1) if $\alpha = (ij)$ and $\alpha' = (lm)$ have one index in common. However, the total number of pairs $(t_\alpha, t_{\alpha'})$ with possible nonzero covariance is of order $O(N^3) = O\{(N^*)^{3/2}\}$, and the total number of zero-covariance pairs is of order $O\{(N^*)^2\}$; with similar notation to that in the proof of Theorem 1, it can be shown that

$$E_p\{V_\xi(T_1)\} = O\{(N^*)^{-1}\}, \quad E_p\{V_\xi(T_2)\} = O\{(n^*)^{-1}(N^*)^{-\frac{1}{2}}\},$$

where $n^* = n(n-1)/2$. The rest of the proof follows directly from that of Theorem 1. □

The pseudo empirical loglikelihood function can also be modified to accommodate all the pairs (ij) using the d_{ij}'s. Let

$$l^*(p) = \sum_{i \in s} \sum_{j>i} d_{ij} \log p_{ij}.$$

The model-calibrated pseudo empirical maximum likelihood estimator of Q is defined as

$$\hat{Q}_{\mathrm{ME}} = N^* \sum_{i \in s} \sum_{j>i} \hat{p}_{ij} \phi(y_i, y_j),$$

where the $\hat{p}_{ij}$'s maximise $l^*(p)$ subject to $p_{ij} > 0$ and

$$\sum_{i \in s} \sum_{j>i} p_{ij} = 1, \quad \sum_{i \in s} \sum_{j>i} p_{ij} E_\xi\{\phi(y_i, y_j) | x_i, x_j\} = \frac{1}{N^*} \sum_{i=1}^{N} \sum_{j=i+1}^{N} E_\xi\{\phi(y_i, y_j) | x_i, x_j\}. \tag{4·2}$$

A theorem that is parallel to Theorem 2 regarding the optimality of $\hat{Q}_{\mathrm{ME}}$ can be similarly established. As usual, any model parameter appearing in constraints (4·1) or (4·2) will be replaced by sample-based design-consistent estimators.

4·2. *Estimation of the population variance*

Note that the population variance can be rewritten as

$$S^2 = \{N(N-1)\}^{-1} \sum_{i=1}^{N} \sum_{j=i+1}^{N} (y_i - y_j)^2.$$

Under model (2·1),

$$E_\xi\{(y_i - y_j)^2 | x_i, x_j\} = \{\mu(x_i, \theta) - \mu(x_j, \theta)\}^2 + \sigma^2\{v^2(x_i) + v^2(x_j)\},$$

and this should be used in constraints (4·1) for optimal estimation. One can also replace (4·1) by two equations,

$$\sum_{i \in s} \sum_{j>i} w_{ij}\{\mu(x_i, \theta) - \mu(x_j, \theta)\}^2 = \sum_{i=1}^{N} \sum_{j=i+1}^{N} \{\mu(x_i, \theta) - \mu(x_j, \theta)\}^2, \tag{4·3}$$

$$\sum_{i \in s} \sum_{j>i} w_{ij}\{v^2(x_i) + v^2(x_j)\} = \sum_{i=1}^{N} \sum_{j=i+1}^{N} \{v^2(x_i) + v^2(x_j)\}, \tag{4·4}$$

to avoid the estimation of σ^2. In many applications $v(x_i) \equiv 1$, in which case the second calibration equation (4·4) becomes $\sum_{i \in s} \sum_{j>i} w_{ij} = N^*$. The resulting estimator $\hat{S}^2_{\mathrm{MC}}$ reduces to the one proposed by Sitter & Wu (2002). Under a linear working model where $\mu(x_i, \theta) = x_i'\theta$, this estimator has a neat form of $\hat{S}^2_{\mathrm{MC}} = \hat{S}^2_{\mathrm{HT}} + \hat{\theta}'(S_x^2 - s_x^2)\hat{\theta}\hat{B}$, where

$$\hat{S}^2_{\mathrm{HT}} = \{N(N-1)\}^{-1} \sum_{i \in s} \sum_{j>i} d_{ij}(y_i - y_j)^2, \quad S_x^2 = (N-1)^{-1} \sum_{i=1}^{N} (x_i - \bar{X})(x_i - \bar{X})',$$

$$s_x^2 = \{N(N-1)\}^{-1} \sum_{i \in s} \sum_{j>i} d_{ij}(x_i - x_j)(x_i - x_j)',$$

and $\hat{B}$ is the estimated regression coefficient of regressing $v_{ij}=(y_i-y_j)^2$ over $u_{ij}=\hat{\theta}'(x_i-x_i)(x_i-x_j)'\hat{\theta}$.

The model-calibrated pseudo empirical maximum likelihood estimator is more useful in this context, because of the normalised positive weights. Note that under model (2·1) constraints (4·2) can be replaced by

$$\sum_{i\in s}\sum_{j>i} p_{ij}=1 \quad (p_{ij}>0), \tag{4·5}$$

$$\sum_{i\in s}\sum_{j>i} p_{ij}\{\mu(x_i,\theta)-\mu(x_j,\theta)\}^2=\frac{1}{N^*}\sum_{i=1}^{N}\sum_{j=i+1}^{N}\{\mu(x_i,\theta)-\mu(x_j,\theta)\}^2, \tag{4·6}$$

$$\sum_{i\in s}\sum_{j>i} p_{ij}\{v^2(x_i)+v^2(x_j)\}=\frac{1}{N^*}\sum_{i=1}^{N}\sum_{j=i+1}^{N}\{v^2(x_i)+v^2(x_j)\}. \tag{4·7}$$

When $v(x_i)\equiv 1$, equation (4·7) reduces to (4·5), and the resulting estimator $\hat{S}^2_{\mathrm{ME}}$ also reduces to the one proposed by Sitter & Wu (2002). A simple and stable algorithm for computing the weights $\hat{p}_{ij}$ is described in Sitter & Wu (2002). Since $\hat{p}_{ij}>0$, the model-calibrated pseudo empirical maximum likelihood estimator is always positive, which is desirable for practical applications.

Under a model with nonhomogeneous variance, including constraint (4·4) or (4·7) will usually improve the efficiency of the resulting estimators, as shown by the simulation study reported in § 5.

4·3. *Variance estimation for the generalised regression estimator*

The generalised regression estimator for the population total or mean is one of the most popular techniques for using auxiliary information from surveys. If the totals X are assumed known, the generalised regression estimator for the population total Y is computed as $\hat{Y}_{\mathrm{GR}}=\hat{Y}_{\mathrm{HT}}+(X-\hat{X}_{\mathrm{HT}})'\hat{\theta}$, where $\hat{Y}_{\mathrm{HT}}=\sum_{i\in s}d_i y_i$ and $\hat{X}_{\mathrm{HT}}=\sum_{i\in s}d_i x_i$ are the conventional Horvitz–Thompson estimators, and $\hat{\theta}$ is the estimated regression coefficient of y over x. Its asymptotic design variance is given by

$$AV_p(\hat{Y}_{\mathrm{GR}})=\sum_{i=1}^{N}\sum_{j=i+1}^{N}(\pi_i\pi_j-\pi_{ij})\left(\frac{e_i}{\pi_i}-\frac{e_j}{\pi_j}\right)^2,$$

where $e_i=y_i-x_i'\theta_N$ and θ_N is the finite population regression coefficient that is estimated by $\hat{\theta}$.

The question of when and how auxiliary information can be used for both the estimation of the population total using a generalised regression estimator and the estimation of its variance can now be answered clearly under the optimal model-calibration approach. Note that $AV_p(\hat{Y}_{\mathrm{GR}})$ has the form of Q with $\phi(y_i,y_j)=(\pi_i\pi_j-\pi_{ij})(e_i/\pi_i-e_j/\pi_j)^2$. Under model (3·1), which is the one that motivated the generalised regression estimator, the optimal calibration variable that should be used in (4·1) is

$$E_\xi\{\phi(y_i,y_j)\,|\,x_i,x_j\}\doteq(\pi_i\pi_j-\pi_{ij})\left\{\frac{v^2(x_i)}{\pi_i^2}+\frac{v^2(x_j)}{\pi_j^2}\right\}\sigma^2.$$

Here we have used the fact that $E_\xi(e_i)=0$. It is now clear that, if model (3·1) has a homogeneous variance structure, that is $v(x_i)\equiv 1$, the calibration variable will be independent of the x_i's. The same auxiliary information cannot be used to improve variance estimation for the generalised regression estimator. On the other hand, under a linear

regression model with nonhomogeneous variance, there will be room for improvement. The constraint that should be used to construct the model-calibration estimator is given by

$$\sum_{i\in s}\sum_{j>i} w_{ij}(\pi_i\pi_j-\pi_{ij})\left\{\frac{v^2(x_i)}{\pi_i^2}+\frac{v^2(x_j)}{\pi_j^2}\right\}=\sum_{i=1}^{N}\sum_{j=i+1}^{N}(\pi_i\pi_j-\pi_{ij})\left\{\frac{v^2(x_i)}{\pi_i^2}+\frac{v^2(x_j)}{\pi_j^2}\right\}.$$

A similar constraint should be used when one estimates $AV_p(\hat{Y}_{\mathrm{GR}})$ using the model-calibrated pseudo empirical likelihood method.

5. Simulation study

In this section, we investigate the optimality of the proposed estimators under the true model and their robustness against model misspecifications. We also compare our estimators to the leading existing competitors.

In the simulation, a fixed finite population of size $N=2000$ was generated from a superpopulation model, ξ. Two superpopulation models were used: model P_1 is a linear regression model with a nonhomogeneous variance structure, so that $y_i=\beta_0+\beta_1x_{1i}+\beta_2x_{2i}+x_{1i}\varepsilon_i$, and model P_2 is the log-linear model $\log y_i=\beta_0+\beta_1x_{1i}+\beta_2x_{2i}+\varepsilon_i$, for $i=1,\ldots,N$. Let $x_i=(1,x_{1i},x_{2i})'$ and $\beta=(\beta_0,\beta_1,\beta_2)'$. The mean function and the variance function for the linear model P_1 are $\mu_i=\mu(x_i,\beta)=x_i'\beta$ and $v_i=v(x_i)=x_{1i}$, and for the nonlinear model P_2 they are $\mu_i=\exp(x_i'\beta)$ and $v_i=\mu_i$, that is $V_\xi(y_i|x_i)=\sigma^2\mu_i^2$. Under the log-linear model P_2, the regularity condition (ii) of Wu & Sitter (2001) would require $N^{-1}\sum_{i=1}^{N}\exp(x_i'\beta)=O(1)$, so certain heavy-tailed distributions such as the log-normal distribution or the gamma distribution with a large scale parameter cannot be used to generate the x values. With this in mind, we generated the x_{1i}'s from a gamma distribution with shape parameter 1 and scale parameter 0·5, and the x_{2i}'s from $|Z|$ where $Z\sim N(0,1)$. Both auxiliary variables take nonnegative values and are skewed to the right, which is quite common in real applications. The values of β_1 and β_2 were both conveniently set to be 1 and β_0 was chosen such that $y_i>0$ for model P_1. The ε_i's are independent and identically distributed as $N(0,\sigma_0^2)$. Four different values of σ_0^2 were used such that the finite population correlation coefficient ρ between y_i and $x_{1i}+x_{2i}$ for model P_1 or between $\log y_i$ and $x_{1i}+x_{2i}$ for model P_2 are 0·9, 0·8, 0·7 and 0·6, respectively. In each simulation run, a simple random sample of size $n=100$ was taken from the finite population, the model parameters $(\beta_0,\beta_1,\beta_2)$ and σ^2 were estimated by the weighted least squares method for model P_1 and by the quasi maximum likelihood method for model P_2, and various estimators were computed. The process was repeated $B=1000$ times. The simulation was programmed in R/S-Plus and the source codes are available from the author upon request.

For the population mean $\bar{Y}$, the conventional calibration estimator of Deville & Särndal (1992) is equivalent to the model-calibration estimator under a linear model. We denote the model-calibration estimator, MC, and the model-calibrated pseudo empirical maximum likelihood estimator, ME, under the linear model P_1 by $\tilde{Y}_{\mathrm{MC}_1}$ and $\tilde{Y}_{\mathrm{ME}_1}$; the corresponding estimators under the log-linear model P_2 are denoted by $\tilde{Y}_{\mathrm{MC}_2}$ and $\tilde{Y}_{\mathrm{ME}_2}$. The performance of an estimator $\tilde{Y}$ is evaluated using its relative bias, RB, and relative efficiency, RE, defined by

$$\mathrm{RB}=B^{-1}\sum_{b=1}^{B}(\tilde{Y}_b-\bar{Y})/\bar{Y},\quad \mathrm{RE}=\mathrm{MSE}(\tilde{Y}_{\mathrm{HT}})/\mathrm{MSE}(\tilde{Y}),$$

where $\mathrm{MSE}(\tilde{Y})=B^{-1}\sum_{b=1}^{B}(\tilde{Y}_b-\bar{Y})^2$ and $\tilde{Y}_b$ is computed from the bth simulated sample.

The Horvitz–Thompson estimator, $\tilde{Y}_{HT}$, is used for baseline comparison. Large values of RE (>1) represent high efficiency.

The simulated absolute values of relative biases are all less than 0·5%. The simulated relative efficiencies are reported in Table 1. Note that $\tilde{Y}_{MC_1}$ and $\tilde{Y}_{ME_1}$ are optimal under model P_1 while $\tilde{Y}_{MC_2}$ and $\tilde{Y}_{ME_2}$ are optimal under model P_2, and other cases are associated with model misspecification. The optimality of the MC and the ME estimators under the true model and the robustness of these estimators against misspecified models are clearly supported by the simulation results. The fact that fitting a log-linear model under P_1 is less serious than fitting a linear model under P_2 is also shown from the simulation. For example, under P_1 with $\rho(y_i, x_{1i}+x_{2i})=0{\cdot}80$, we found that $\rho(\log y_i, x_{1i}+x_{2i})=0{\cdot}77$, while, under P_2 with $\rho(\log y_i, x_{1i}+x_{2i})=0{\cdot}80$, we observed $\rho(y_i, x_{1i}+x_{2i})=0{\cdot}59$. Under all cases the ME estimator and the MC estimator perform similarly to each other.

Table 1. *Simulated relative efficiencies of estimators for the population mean relative to the Horvitz–Thompson estimator*

ρ	$\tilde{Y}_{MC_1}$	$\tilde{Y}_{ME_1}$	$\tilde{Y}_{MC_2}$	$\tilde{Y}_{ME_2}$	$\tilde{Y}_{MC_1}$	$\tilde{Y}_{ME_1}$	$\tilde{Y}_{MC_2}$	$\tilde{Y}_{ME_2}$
	Model P_1				Model P_2			
0·60	1·62	1·61	1·58	1·57	1·01	1·02	1·08	1·05
0·70	2·07	2·05	1·96	1·96	1·10	1·11	1·56	1·49
0·80	2·91	2·87	2·52	2·58	1·25	1·28	2·42	2·28
0·90	5·71	5·66	3·87	4·24	1·52	1·58	5·54	5·19

For the distribution function $F_Y(t)$, four estimators are computed under the linear regression model P_1, the model-based estimator of Chambers & Dunstan (1986), CD, the model-assisted difference estimator of Rao et al. (1990), RKM, the optimal model-calibrated empirical likelihood estimator, ME_1, and the pseudo empirical likelihood estimator of Chen & Sitter (1999) calibrated directly over the x variables, CS. The optimal model-calibrated empirical likelihood estimator under the logistic regression model (3·3) is denoted by ME_2. Once again, the Horvitz–Thompson estimator $\hat{F}_{HT}(t)$ is used for baseline comparison. All estimators are computed at five different population quantiles t_α with $\alpha=0{\cdot}10$, 0·30, 0·50, 0·70 and 0·90, and the optimal weights $\hat{p}_i$ are computed using the particular value of t_α.

Table 2 presents the simulated relative efficiencies under populations P_1 and P_2, both with $\rho=0{\cdot}80$. Results for other values of ρ demonstrated a similar pattern, together with reduced relative efficiency for all estimators as ρ decreases. The simulated relative biases are all within 2% except for the Chambers & Dunstan estimator which is outrageously biased under model P_2. Table 2 can be summarised as follows: the optimal estimator ME_1 and the Rao et al. difference estimator show good and similar performance under the true

Table 2. *Simulated efficiencies of estimators for the distribution function relative to the Horvitz–Thompson estimator, at five α-quantiles*

	α					α				
Method	0·10	0·30	0·50	0·70	0·90	0·10	0·30	0·50	0·70	0·90
	Model P_1					Model P_2				
CD	7·85	9·85	8·60	7·96	6·95	0·05	1·13	0·61	0·34	0·26
RKM	2·04	2·75	2·56	2·67	2·28	0·93	1·28	1·40	1·45	1·32
ME_1	1·99	2·75	2·56	2·65	2·17	1·07	1·29	1·40	1·45	1·31
ME_2	1·56	2·13	2·15	2·31	1·85	1·14	1·48	1·66	1·96	1·62
CS	1·08	1·47	1·76	1·93	1·46	1·11	1·35	1·60	1·83	1·42

model P_1 and are robust against the misspecified model P_2; the model-based Chambers & Dunstan estimator has superb performance under the true model P_1 but totally collapsed under the misspecified model P_2; the estimator ME_2 based on an assumed logistic regression model performs reasonably well. It is less efficient than ME_1 under the true model but has better performance when the model is misspecified as P_2; the Chen & Sitter estimator shows marginal to moderate gain of efficiency at all cases.

The optimal estimator ME_1 may be preferred to the Rao et al. estimator for real applications. Let $g_i = g(x_i, t)$ be given by (3·2). The Rao et al. estimator is

$$\hat{F}_{\text{RKM}}(t) = \hat{F}_{\text{HT}}(t) + N^{-1}\left(\sum_{i=1}^{N} g_i - \sum_{i \in s} d_i g_i\right).$$

It is not a distribution function and can indeed take values outside $[0, 1]$. On the other hand, the ME_1 estimator is $\hat{F}_{\text{ME}_1}(t) = \sum_{i \in s} p_i I(y_i \leqslant t)$ with $p_i > 0$ and $\sum_{i \in s} p_i = 1$. Quantile estimation can be easily achieved through direct inversion of $\hat{F}_{\text{ME}_1}(t)$. Theoretically we can show that the ME_1 estimator will perform at least as well as the Rao et al. estimator under large samples (Chen & Wu, 2002).

For the population variance S^2, the calibration estimator and the empirical maximum likelihood estimator of Sitter & Wu (2002) using the single constraint (4·3) or (4·6) are denoted by $\hat{S}^2_{\text{MC}_1}$ and $\hat{S}^2_{\text{ME}_1}$, respectively. The optimal estimators using both constraints (4·3) and (4·4), or (4·6) and (4·7), are denoted by $\hat{S}^2_{\text{MC}_2}$ and $\hat{S}^2_{\text{ME}_2}$. The uniform weights $q_{ij} = 1$ are used for the MC estimators. These estimators are computed based on model P_1. The relative bias and relative efficiency are similarly defined and comparisons are made with the baseline Horvitz–Thompson estimator $\hat{S}^2_{\text{HT}}$.

The simulated relative efficiencies of all four estimators under the true model are reported in Table 3. The absolute values of the simulated relative bias are all less than 4%. The estimators $\hat{S}^2_{\text{MC}_1}$ and $\hat{S}^2_{\text{ME}_1}$ perform well when the variable of the mean function, $\mu(x_i, \hat{\beta}) = x_i'\hat{\beta}$, is a strong predictor of the response variable, corresponding to a high value of ρ, but they deteriorate quickly as the relationship becomes weak. The optimal estimators and $\hat{S}^2_{\text{MC}_2}$ and $\hat{S}^2_{\text{ME}_2}$, which use auxiliary information from both the mean function $\mu(x_i, \hat{\beta})$ and the variance function $v(x_i)$, perform well for all cases, and their loss of efficiency when ρ is reduced is less dramatic.

Table 3. *Simulated efficiencies of estimators for the population variance relative to the Horvitz–Thompson estimator*

ρ	$\hat{S}^2_{\text{MC}_1}$	$\hat{S}^2_{\text{ME}_1}$	$\hat{S}^2_{\text{MC}_2}$	$\hat{S}^2_{\text{ME}_2}$
0·60	1·72	1·68	2·20	2·05
0·70	1·82	1·77	2·25	2·13
0·80	2·00	1·94	2·34	2·25
0·90	2·55	2·47	2·77	2·68

6. DISCUSSION

The proposed optimal calibration approach requires specification of a mean function $\mu(x_i, \theta)$ and/or a variance function $v(x_i)$ from the model. A general discussion of model building and diagnostics using complex survey data is beyond the scope of this paper and requires further research. In many applications, the parametric linear regression model (3·1) will probably be used. In the case of a single x variable, Breidt & Opsomer (2000)

used nonparametric smoothing technique to find the model expectations of the response variable. Extending the method to multiple x variables seems possible.

An important feature of the results presented in this paper is that the optimality of the model-calibration or the model-calibrated pseudo empirical maximum likelihood estimators is independent of the sampling design as long as the latter is 'regular'. This is in contrast to the results of Godambe & Thompson (1973), or Cassel et al. (1976), where an optimal estimator corresponds to a particular sampling design. The independence of the optimality of an estimator to sampling design is practically appealing when such an estimator is to be constructed at the estimation stage. Some fundamental issues in using auxiliary information from surveys can now be addressed more clearly in the light of this optimal calibration approach.

(i) The effective use of auxiliary information from survey data depends on both the population quantities to be estimated and the actual relationship between the response variable and the covariates. Blindly calibrating over auxiliary variables is usually not a good approach.

(ii) The benchmark constraints used in (1·1) are justifiable if the relationship between y and x is close to linear and the parameter of interest is the population mean or total. In this case the resulting conventional calibration estimator of $\bar{Y}$ is identical to the optimal model-calibration estimator obtained using $\hat{\mu}_i = x_i'\hat{\theta}$ as the calibration variable, so benchmarking implies efficient estimation.

(iii) If the relationship between y and x is linear, knowing $\bar{X}$ is 'sufficient' for efficient estimation of the population mean $\bar{Y}$ or the total Y. If the relationship is nonlinear, or the parameters of interest involve a nonlinear function, complete auxiliary information and/or more advanced modelling are essential for 'optimal' estimation.

(iv) The variance function $v(x_i)$ from model (2·1) does not play a role in the construction of optimal calibration estimators for the population mean or total. However, this is not the case for the optimal estimation of the finite population distribution function, the population variance or other second-order population quantities where $v(x_i)$ is equally as important as the mean function $\mu(x_i, \theta)$.

(v) Auxiliary information can sometimes be triply used at the design stage, the estimation of the population mean or total using a generalised regression estimator, and the estimation of its variance through calibration. Such situations can be identified under the optimal calibration approach.

For cases where complete auxiliary information is required for optimal estimation but such information is not available, the optimal calibration approach can be combined with two-phase sampling in which the large first-phase sample measured over the covariates is treated as 'complete' auxiliary information; see Wu & Luan (2003) for further details.

Acknowledgement

This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada. The author thanks Professor Randy R. Sitter, Dr Steve Drekic, the editor, associate editor and two referees for helpful comments and suggestions that greatly improved the paper.

Appendix

Proofs of Theorems 1 *and* 2

Proof of Theorem 1. Without loss of generality, we consider the chi-squared distance measure with the weights q_i satisfying $N^{-1}\sum_{i=1}^{N} q_i^6 = O(1)$ and $q_i \geqslant q$ for some constant $q > 0$. It can be

easily shown that minimising Φ_s subject to (1·2) leads to

$$\tilde{Y}_{C_u} = \frac{1}{N}\sum_{i\in s} d_i y_i + \frac{1}{N}\left(\sum_{i=1}^{N} u_i - \sum_{i\in s} d_i u_i\right)\hat{B},$$

where $u_i = u(x_i)$ and $\hat{B} = (\sum_{i\in s} d_i q_i u_i y_i)/(\sum_{i\in s} d_i q_i u_i^2)$. Under a regular sampling design, $AV_p(\tilde{Y}_{C_u}) = V_p(T)$, where

$$T = \frac{1}{N}\sum_{i\in s} d_i y_i + \frac{1}{N}\left(\sum_{i=1}^{N} u_i - \sum_{i\in s} d_i u_i\right)B_N$$

and $B_N = (\sum_{i=1}^{N} u_i q_i y_i)/(\sum_{i=1}^{N} q_i u_i^2)$.

Let $\mu_i = \mu(x_i, \theta)$, $\bar{\mu} = E_\xi(\bar{Y}) = N^{-1}\sum_{i=1}^{N}\mu_i$ and $B_\xi(T) = E_\xi(T) - \bar{\mu}$. Since $E_p(T) = \bar{Y}$ and $V_p(T) = E_p(T - \bar{Y})^2$, it is straightforward to show that

$$E_\xi\{V_p(T)\} = E_p\{V_\xi(T)\} + E_p\{B_\xi(T)\}^2 - V_\xi(\bar{Y}).$$

Note that E_ξ and V_ξ are conditional on the given x_i's. Let

$$U^2 = N^{-1}\sum_{i=1}^{N} q_i u_i^2, \quad D = N^{-1}\left(\sum_{i=1}^{N} u_i - \sum_{i\in s} d_i u_i\right).$$

We can rewrite T as $T_1 + T_2$, where $T_1 = N^{-1}\sum_{i\in s} d_i y_i$ and $T_2 = DU^{-2}N^{-1}\sum_{i=1}^{N} q_i u_i y_i$. We have

$$E_p\{V_\xi(T)\} = E_p\{V_\xi(T_1)\} + E_p\{V_\xi(T_2)\} + 2E_p\{\mathrm{cov}_\xi(T_1, T_2)\},$$

where $\mathrm{cov}_\xi(T_1, T_2)$ denotes the covariance under the model. It can be seen that

$$E_p\{V_\xi(T_1)\} = \frac{1}{N^2}\sum_{i=1}^{N} d_i v^2(x_i)\sigma^2 = O\left(\frac{1}{n}\right), \tag{A·1}$$

$$E_p\{V_\xi(T_2)\} = \{E_p(D^2)\}U^{-4}\frac{1}{N^2}\sum_{i=1}^{N} q_i^2 u_i^2 v^2(x_i)\sigma^2 = O\left(\frac{1}{nN}\right). \tag{A·2}$$

Here the last step in (A·1) follows from the Condition 1 that $\max_{i\in s} nd_i/N = O(1)$ and the assumption that $\{v(x_1), v(x_2), \ldots\} \in L$, and the last step in (A·2) follows by noting that

$$E_p(D^2) = V_p\left(N^{-1}\sum_{i\in s} d_i u_i\right) = O(n^{-1}),$$

U^2 is bounded from zero, and $N^{-1}\sum_{i=1}^{N} q_i^2 u_i^2 v^2(x_i) = O(1)$ under the assumed finite moment conditions.

It also follows from $|\mathrm{cov}_\xi(T_1, T_2)| \leqslant \{V_\xi(T_1)\}^{1/2}\{V_\xi(T_2)\}^{1/2}$ that

$$\{E_p|\mathrm{cov}_\xi(T_1, T_2)|\}^2 \leqslant E_p\{V_\xi(T_1)\}E_p\{V_\xi(T_2)\},$$

which implies that $E_p\{\mathrm{cov}_\xi(T_1, T_2)\} = O(n^{-3/2})$. When n is large, the leading term in $E_p\{V_\xi(T)\}$ is $E_p\{V_\xi(T_1)\}$, which is independent of the choice of sequence C_u. The term $V_\xi(\bar{Y})$ is also independent of C_u.

For the term $E_p\{B_\xi(T)\}^2$, note that

$$B_\xi(T) = \frac{1}{N}\sum_{i\in s} d_i(\mu_i - u_i B) - \frac{1}{N}\sum_{i=1}^{N}(\mu_i - u_i B),$$

where $B = \sum_{i=1}^{N} q_i u_i \mu_i / \sum_{i=1}^{N} q_i u_i^2$. It follows that $E_p\{B_\xi(T)\} = 0$ and

$$E_p\{B_\xi(T)\}^2 = V_p\{B_\xi(T)\} = V_p\left\{N^{-1}\sum_{i\in s} d_i(\mu_i - u_i B)\right\} = O(n^{-1}).$$

Minimising $E_\xi\{AV_p(\tilde{Y}_{C_u})\}$ amounts to minimising $E_p\{B_\xi(T)\}^2$. The choice $C_\mu = (\mu_1, \mu_2, \ldots)$ results

in $B=1$ and $E_p\{B_\xi(T)\}^2=0$. □

Proof of Theorem 2. By Theorem 1 of Chen & Sitter (1999), we have

$$\tilde{Y}_{\mathrm{ME}}=\left(\sum_{i\in s}d_i\right)^{-1}\sum_{i\in s}d_i y_i+\left\{\frac{1}{N}\sum_{i=1}^{N}u(x_i)-\left(\sum_{i\in s}d_i\right)^{-1}\sum_{i\in s}d_i u(x_i)\right\}\hat{B}+o_p(n^{-1/2}),$$

where $\hat{B}$ is similarly defined as in Theorem 1 with $q_i=1$.

The term $T_1^*=(\sum_{i\in s}d_i)^{-1}\sum_{i\in s}d_i y_i$ is a ratio-type estimator and its design-based variance $V_p(T_1^*)$ is not the same as $V_p(T_1)$, where $T_1=N^{-1}\sum_{i\in s}d_i y_i$. However, since $\sum_{i\in s}d_i$ is a constant under the superpopulation model, the conclusion about $E_p\{V_\xi(T_1)\}$ in Theorem 1 can also be restated here in terms of T_1^*. The remaining part of the proof is similar to the proof of Theorem 1 and is omitted. □

References

Breidt, F. J. & Opsomer, J. D. (2000). Local polynomial regression estimators in survey sampling. *Ann. Statist.* **28**, 1026–53.

Cassel, C. M., Särndal, C. E. & Wretman, J. H. (1976). Some results on generalised difference estimation and generalized regression estimation for finite populations. *Biometrika* **63**, 615–20.

Chambers, R. L. & Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika* **73**, 597–604.

Chen, J. & 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. & Wu, C. (2002). Estimation of distribution function and quantiles using the model-calibrated pseudo empirical likelihood method. *Statist. Sinica* **12**, 1223–39.

Chen, J., Sitter, R. R. & Wu, C. (2002). Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys. *Biometrika* **89**, 230–7.

Deville, J. C. & Särndal, C. E. (1992). Calibration estimators in survey sampling. *J. Am. Statist. Assoc.* **87**, 376–82.

Godambe, V. P. (1955). A unified theory of sampling from finite populations. *J. R. Statist. Soc.* B **17**, 267–78.

Godambe, V. P. & Thompson, M. E. (1973). Estimation in sampling theory with exchangeable prior distributions. *Ann. Statist.* **1**, 1212–21.

Isaki, C. T. & Fuller, W. A. (1982). Survey design under the regression superpopulation model. *J. Am. Statist. Assoc.* **77**, 89–96.

Rao, J. N. K., Kovar, J. G. & Mantel, H. J. (1990). On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika* **77**, 365–75.

Sitter, R. R. & Wu, C. (2002). Efficient estimation of quadratic finite population functions in the presence of auxiliary information. *J. Am. Statist. Assoc.* **97**, 535–43.

Wu, C. & Luan, Y. (2003). Optimal calibration estimators under two-phase sampling. *J. Offic. Statist.* **19**, 119–31.

Wu, C. & Sitter, R. R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *J. Am. Statist. Assoc.* **96**, 185–93.

[*Received January* 2002. *Revised February* 2003]