The Effective Use of Complete Auxiliary Information Through Model-calibration and Empirical Likelihood

Changbao Wu, University of Waterloo and Randy R. Sitter, Simon Fraser University
Changbao Wu, Dept. of Statistics and Actuarial Science, University of Waterloo
Waterloo, ON N2L 3G1 Canada (cbwu@math.uwaterloo.ca)

Key Words: Distribution function; Empirical likelihood; Estimating equations; Generalized linear models; Generalized regression estimator; Superpopulation.

Abstract:

A unified framework has been attempted to address the problem "How can complete auxiliary information be effectively used from survey data". The proposed model-calibration estimators can effectively handle linear or non-linear models and reduce to the conventional calibration estimators of Deville and Särndal (1992) and/or the pseudo-empirical maximum likelihood estimators of Chen and Sitter (1999) under linear models.

1 Introduction

In sample surveys, auxiliary information on the finite population is often used to increase the precision of estimators of the population mean, total or distribution function. In the simplest settings, ratio and regression estimators incorporate known finite population means of auxiliary variables. For more general situations, there have been three main methods proposed in the literature which can be categorized as model-assisted approaches: the generalized regression estimator (GR) (Cassel, Särndal and Wretman, 1976; Särndal, 1980); calibration estimators (Deville and Särndal, 1992); and more recently empirical likelihood methods (Chen and Qin, 1993; Zhong and Rao, 1998; Chen and Sitter, 1999). All of these methods have only been discussed in the context of a linear regression working model and essentially incorporate the auxiliary variables through their known population means even when the auxiliary variables are known for every unit in the population.

In this paper, we consider the use of more complex working models in obtaining model-assisted estimators by generalizing the calibration method above. We term the approach model-calibration for reasons which will become readily apparent. We argue that, under a general modeling process, complete auxiliary information should be incorporated into the construction of estimators through fitted values. How to do this properly is fairly straightforward in the case of a GR (see Section 3) but not so for calibration. We introduce a general framework from which to do this that is simple, and reduces to the usual estimators under a linear model.

Once this generalization is realized, some interesting relationships between a linear model and the use of complete auxiliary information become more obvious and are discussed. Also, some differences between the approaches become more distinct. For example it has been noted that the calibration estimator reduces to a GR under a chi-square distance measure (Deville and Särndal, 1992), where an underlying linear regression model is used. This is no longer the case when the methods are generalized to nonlinear models, and the proposed model-calibration method performs better.

In Section 2 we briefly review the calibration method and discuss its implicit model-assisted nature and relationship to a linear model. In Section 3 we propose a model-calibration method for incorporating auxiliary information into estimation of the population mean under a very general model which includes linear and nonlinear regression and generalized linear models as special cases. We go on to show that the resulting estimator is asymptotically design-unbiased and reduces to the usual calibration method under a linear regression model. Also in Section 3, we discuss the extension of the GR and the pseudo-empirical maximum likelihood (EL) methods to the general model and demonstrate that unlike in the linear model case, the extended calibration and the extended GR do not yield the same estimator. We go on to demonstrate, through a small simulation study, that the model-calibration estimator and the EL are superior.

This research was supported by a grant from the Natural Sciences and Engineering Council of Canada.

2 How the Usual Calibration Method Relates to a Linear Model

Consider a finite population consisting of N identifiable units. Associated with the i-th unit are, the study variable, y_i, and a vector of auxiliary variables, $\boldsymbol{x}_i$. The values $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N$ are known for the entire population but y_i is known only if the i-th unit is selected in the sample, s. Assume the inclusion probabilities $\pi_i = \Pr(i \in s)$ are strictly positive. For the moment we will restrict attention to estimating the population total $Y = \sum_{i=1}^{N} y_i$.

Deville and Särndal (1992) introduce the notion of a calibration estimator of Y which is constructed as $\hat{Y}_C = \sum_{i \in s} w_i y_i$, where the calibration weights w_i's are chosen to minimize their average distance Φ_s from the basic design weights, $d_i = 1/\pi_i$, that are used in the Horvitz-Thompson estimator, $\hat{Y}_{HT} = \sum_{i \in s} d_i y_i$; most commonly

$$\Phi_s = \sum_{i \in s} (w_i - d_i)^2/(d_i q_i), \qquad (1)$$

subject to the constraint

$$\sum_{i \in s} w_i \boldsymbol{x}_i = \boldsymbol{X}, \qquad (2)$$

where the q_i's are known positive weights unrelated to d_i. The resulting calibration estimator is

$$\hat{Y}_C = \sum_{i \in s} w_i y_i = \hat{Y}_{HT} + (\boldsymbol{X} - \hat{\boldsymbol{X}}_{HT})' \hat{\boldsymbol{B}}, \qquad (3)$$

where $\hat{\boldsymbol{X}}_{HT} = \sum_{i \in s} d_i \boldsymbol{x}_i$ and $\hat{\boldsymbol{B}} = \{\sum_{i \in s} d_i q_i \boldsymbol{x}_i \boldsymbol{x}_i'\}^{-1} \sum_{i \in s} d_i q_i \boldsymbol{x}_i y_i$. The uniform weights $q_i = 1$ are used in most applications, but unequal weights can also be motivated as in example 1 of Deville and Särndal (1992). The calibrated weights, w_i, give perfect estimates when applied to the auxiliary variables. Deville and Särndal (1992) argue that "the weights that perform well for the auxiliary variable also should perform well for the study variable". However, it is an implicit underlying assumption that y and $\boldsymbol{x}$ are linearly related that makes this a valid argument. For example, in the case of scaler x and $\boldsymbol{x}_i' = (1, x_i)$ is used in (2), it is clear that $y_i = \beta_0 + \beta_1 x_i$ implies $\hat{Y}_C = Y$. If a curved relationship exists between y and x, the so constructed calibration estimator could be very inefficient. For instance, if $\log(y_i) \doteq \beta_0 + \beta_1 x_i$, then there is no compelling reason to use $\hat{Y}_C$.

The point we want to illustrate is that, it is the model structure (relationship between y and $\boldsymbol{x}$) that determines how the auxiliary information should best be used. In fact, Deville and Särndal (1992) show that, for any Φ_s, $\hat{Y}_C$ is asymptotically equivalent to (3), which is the generalized regression estimator, $\hat{Y}_{GR}$, and the GR is motivated as a model-assisted estimator using a linear model (Särndal, 1980). Another point relates to the issue of complete information on the $\boldsymbol{x}$ variables (i.e. known for all units in the population) versus only knowing the value of their population totals, $\boldsymbol{X}$. The GR is motivated by using the predicted values from a linear model for each $\boldsymbol{x}_i$. However, the resulting estimator in (3) only needs $\boldsymbol{X}$ to be implemented. As we will see, this is due to the use of a linear model.

3 Model-calibration Estimator of the Mean

We will use a model-assisted approach. That is, our estimator of $\bar{Y}$ will be design-consistent but will be particularly efficient under a working model. This can be accomplished by first using the $(y_i, \boldsymbol{x}_i)$ for $i \in s$ to build the model and then calibrating to the predicted values from the model using: i) a direct calibration argument such as was discussed in the previous section; ii) using a pseudo-empirical likelihood approach (Chen and Sitter, 1999); or iii) using a generalized difference estimator (Cassel, Särndal and Wretman, 1976; Särndal, 1980). We will briefly discuss the modeling step first and then consider these three methods of calibrating on the predicted values.

3.1 Modeling

Assume the relationship between y and $\boldsymbol{x}$ can be described by a superpopulation model through the first and second moments,

$$E_\xi(y_i|\boldsymbol{x}_i) = \mu(\boldsymbol{x}_i, \boldsymbol{\theta}),\ V_\xi(y_i|\boldsymbol{x}_i) = v_i^2 \sigma^2,$$

$$i = 1, 2, \cdots, N, \qquad (4)$$

where $\boldsymbol{\theta} = (\theta_0, ..., \theta_p)'$ and σ^2 are unknown superpopulation parameters, $\mu(\boldsymbol{x}, \boldsymbol{\theta})$ is a known function of $\boldsymbol{x}$ and $\boldsymbol{\theta}$, the v_i's are known constants for given $\boldsymbol{x}_i$'s or$\mu_i = \mu(\boldsymbol{x}_i, \boldsymbol{\theta})$, and E_ξ and V_ξ denote the expectation and variance with respect to the superpopulation model. We also assume that $(y_1, \boldsymbol{x}_1), \cdots, (y_N, \boldsymbol{x}_N)$ are mutually independent.

The model structure (4) is quite general and includes two very important cases: (i) the linear or non-linear regression model,

$$y_i = \mu(\boldsymbol{x}_i, \boldsymbol{\theta}) + v_i \varepsilon_i, \quad i = 1, 2, \cdots, N, \qquad (5)$$

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

$$g(\mu_i) = \boldsymbol{x}_j'\boldsymbol{\theta},\ V_\xi(y_i|\boldsymbol{x}_i) = v(\mu_i),\ \ i = 1, 2, \cdots, N, \tag{6}$$

where $\mu_i = E_\xi(y_i|\boldsymbol{x}_i)$, $g(\cdot)$ is a link function and $v(\cdot)$ is a variance function.

Consider a design-based method for estimating the model parameters. When a model-based approach is employed, $(y_i, \boldsymbol{x}_i), i \in s$ is viewed as an *iid* sample from the superpopulation. The superpopulation parameters, $\boldsymbol{\theta}$, can then be estimated using standard procedures. Under the design-based framework, the sample data may not follow the same model structure as that of the whole finite population under a complex sampling scheme and $\boldsymbol{\theta}$ may be meaningless from the design-based point of view. In this case, $\boldsymbol{\theta}$ is replaced by $\boldsymbol{\theta}_N$, an estimate of $\boldsymbol{\theta}$ based on the data from the entire population. $\boldsymbol{\theta}_N$ is then estimated by $\hat{\boldsymbol{\theta}}$, a design-based estimate from the sampled data (Godambe and Thompson, 1986).

For illustration, consider two important cases.

Case I. $\boldsymbol{\theta}_N$ can be expressed explicitly as functions of population totals for properly defined population variables. For example, under a linear regression model, $\boldsymbol{\theta}_N$ is the regression parameter of the finite population: $\boldsymbol{\theta}_N = (\boldsymbol{X}_N'\boldsymbol{X}_N)^{-1}\boldsymbol{X}_N'\boldsymbol{y}_N$, where $\boldsymbol{X}_N$ is the $N \times (p+1)$ matrix with rows $(1, \boldsymbol{x}_i')$ for $i = 1, ..., N$ and $\boldsymbol{y}_N = (y_1, ..., y_N)'$. A design-based estimator $\hat{\boldsymbol{\theta}}$ is obtained by plugging in design-based estimates for various population totals in $\boldsymbol{\theta}_N$: $\hat{\boldsymbol{\theta}} = (\boldsymbol{X}_n'\boldsymbol{\Pi}^{-1}\boldsymbol{X}_n)^{-1}\boldsymbol{X}_n'\boldsymbol{\Pi}^{-1}\boldsymbol{y}_n$, where $\boldsymbol{\Pi} = diag(\pi_1, \cdots, \pi_n)$ and $\boldsymbol{X}_n$ and $\boldsymbol{y}_n$ in obvious notation.

Case II. $\boldsymbol{\theta}_N$ is defined by estimating equations. Suppose that the generalized linear model (6) is assumed. We define $\boldsymbol{\theta}_N$ as the maximum quasi-likelihood estimator of $\boldsymbol{\theta}$ based on the entire finite population, i.e., the solution of the estimating equation (Molina and Skinner, 1992):

$$\sum_{i=1}^{N} \boldsymbol{X}_i[g'\{\mu(\boldsymbol{x}_i, \boldsymbol{\theta})\}v\{\mu(\boldsymbol{x}_i, \boldsymbol{\theta})\}]^{-1}[y_i - \mu(\boldsymbol{x}_i, \boldsymbol{\theta})] = 0\,, \tag{7}$$

where $\boldsymbol{X}_i' = (1, \boldsymbol{x}_i')$ and $g'(u) = dg(u)/du$. The estimating function on the left hand side of (7) is a population total, $\hat{\boldsymbol{\theta}}$ is defined as the solution of the design-based sample version of (7), i.e., the solution of the following estimating equation:

$$\sum_{i\in s} d_i\boldsymbol{X}_i[g'\{\mu(\boldsymbol{x}_i, \boldsymbol{\theta})\}v\{\mu(\boldsymbol{x}_i, \boldsymbol{\theta})\}]^{-1}[y_i - \mu(\boldsymbol{x}_i, \boldsymbol{\theta})] = 0\,.$$

The estimate $\hat{\boldsymbol{\theta}}$ is then obtained by standard Newton-Raphson iterative procedures. Under certain regularity conditions (similar to those used by Binder, 1983), it can be shown that in both Cases I and II, $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_N + O_p(n^{-1/2})$.

3.2 Model Calibration

Under model (4), auxiliary information should be used through the fitted values $\mu(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}), i = 1, \cdots, N$. To do this we define the model-calibration estimator of $\bar{Y}$ as $\hat{\bar{Y}}_{MC} = N^{-1}\sum_{i\in s} w_i y_i$, where the calibrated weights, w_i, minimize an average distance between w_i's and d_i's, subject to

$$N^{-1}\sum_{i\in s} w_i = 1\,, \quad \sum_{i\in s} w_i\mu(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}) = \sum_{i=1}^{N} \mu(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}).$$

One should note that in the original formulation of the calibration estimator presented in the previous section, the constraint $N^{-1}\sum_{i\in s} w_i = 1$ is not present. If this constraint is added, the resulting estimator under no auxiliary information is $\hat{\bar{Y}} = \sum_{i\in s} d_i y_i / \sum_{i\in s} d_i$ and not $\hat{\bar{Y}}_{HT} = N^{-1}\sum_{i\in s} d_i y_i$. It was illustrated in Rao (1966) and later in the more well known Basu (1971) elephant example that even though the first estimator estimates the population size N and the second uses its known quantity, the first has better properties. This is true for calibration generally. This constraint arises quite naturally in the case of pseudo-empirical maximum likelihood estimators (Chen and Sitter, 1999).

We will restrict our discussion to the chi-square distance given in (1). The resulting model-calibration estimator then follows directly from the development in Deville and Särndal (1992) by treating the $\hat{\mu}_i = \mu(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}})$ as a scalar auxiliary variable and is given by

$$\hat{\bar{Y}}_{MC} = \hat{\bar{Y}}_{HT} + \{N^{-1}\sum_{i=1}^{N} \hat{\mu}_i - N^{-1}\sum_{i\in s} d_i\hat{\mu}_i\}\hat{B}\ , \tag{8}$$

where $\hat{B} = \sum_{i\in s} d_i q_i(\hat{\mu}_i - \bar{\mu})(y_i - \bar{y}) / \sum_{i\in s} d_i q_i(\hat{\mu}_i - \bar{\mu})^2$, $\bar{y} = \sum_{i\in s} d_i q_i y_i / \sum_{i\in s} d_i q_i$ and $\bar{\mu} = \sum_{i\in s} d_i q_i\hat{\mu}_i / \sum_{i\in s} d_i q_i$.

If constraint $N^{-1}\sum_{i\in s} w_i = 1$ is dropped, the single calibration equation $\sum_{i\in s} w_i\mu(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}) = \sum_{i=1}^{N} \mu(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}})$ yields

$$\hat{\bar{Y}}^*_{MC} = \hat{\bar{Y}}_{HT} + \{N^{-1}\sum_{i=1}^{N} \hat{\mu}_i - N^{-1}\sum_{i\in s} d_i\hat{\mu}_i\}\hat{B}^*\,, \tag{9}$$

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

The important properties of (8) and (9) are summarized in the following theorem. We assume that there is a sequence of sampling designs and a sequence of finite populations, indexed by ν. Both the sample size n_ν and the population size N_ν approach infinity as $\nu \to \infty$. All limiting processes are understood to be as $\nu \to \infty$, but the ν is suppressed to simplify notation.

Theorem 1. *1) Suppose (i) $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_N + O_p(n^{-1/2})$; (ii) $\boldsymbol{\theta}_N \to \boldsymbol{\theta}$; and (iii) for each $\boldsymbol{x}_i$, $\partial\mu(\boldsymbol{x}_i, t)/\partial t$ is continuous in t and $|\partial\mu(\boldsymbol{x}_i, t)/\partial t| \leq h(\boldsymbol{x}_i, \boldsymbol{\theta})$ for t in a neighborhood of $\boldsymbol{\theta}$, and $N^{-1}\sum_{i=1}^N h(\boldsymbol{x}_i, \boldsymbol{\theta}) = O(1)$. Then $\hat{\bar{Y}}_{MC}$ and $\hat{\bar{Y}}^*_{MC}$ are both asymptotically design-unbiased. They are also both approximately model-unbiased under the general model (4) if the design-based estimator $\hat{\boldsymbol{\theta}}$ is close to $\boldsymbol{\theta}$;*

2) If $q_i = 1/v_i^2$ in Φ_s, then both $\hat{\bar{Y}}_{MC}$ and $\hat{\bar{Y}}^*_{MC}$ reduce to the conventional calibration estimator (or GR) under a linear model, where

$$\mu(\boldsymbol{x}_i, \boldsymbol{\theta}) = \theta_0 + \theta_1 x_{i1} + \cdots + \theta_p x_{ip}, \qquad (10)$$

in (5).

Proof: See Wu (1999).

Thus, both $\hat{\bar{Y}}_{MC}$ and $\hat{\bar{Y}}^*_{MC}$ are model-assisted in this sense and can handle any linear or non-linear models. That is, they are both design-consistent irrespective of whether the model holds and particularly efficient under model (4). Also, in the case of no modeling error, i.e. $y_i = \mu_i$, we have $\hat{\bar{Y}}_{MC} = \hat{\bar{Y}}^*_{MC} = \bar{Y}$. In addition, both $\hat{\bar{Y}}_{MC}$ and $\hat{\bar{Y}}^*_{MC}$ reduce to the conventional calibration estimator (Deville and Särndal, 1992) (same as the GR) under a linear model.

3.3 Pseudo Empirical Likelihood Approach

Chen and Sitter (1999) propose using a pseudo-empirical maximum likelihood estimator obtained by maximizing

$$\hat{l}(p) = \sum_{i \in s} d_i \log p_i, \qquad (11)$$

which is a design-unbiased estimator of the log-empirical likelihood one would use if one had the entire population: $E_p(\sum_{i\in s} d_i \log p_i) = \sum_{i=1}^N \log p_i$. Here, E_p refers to expectation under the sampling design. For auxiliary information of the general form $N^{-1}\sum_{i=1}^N u_i = 0$, with $u_i = u(y_i, \boldsymbol{x}_i)$, the method then reduces to maximizing (11) subject to

$$\sum_{i\in s} p_i = 1, \ \sum_{i\in s} p_i u_i = 0 \ \ (0 \leq p_i \leq 1). \qquad (12)$$

Chen and Sitter (1999) then go on to primarily focus on estimating $\bar{Y}$ with $\bar{\boldsymbol{X}}$ known, i.e. $u(y_i, \boldsymbol{x}_i) = (\boldsymbol{x}_i - \bar{\boldsymbol{X}}_N)$. The resulting estimator is asymptotically equivalent to a GR discussed in the next section (Note that for vector valued $\boldsymbol{x}_i - \bar{\boldsymbol{X}}_N$ this needs a vector Lagrange multiplier to solve). Thus, much like the calibration method, there is implicit use of a linear relationship between y and $\boldsymbol{x}$. To extend to model (4) we merely define $u_i = u(y_i, \boldsymbol{x}_i) = \mu(\boldsymbol{x}_i, \boldsymbol{\theta}) - N^{-1}\sum_{i=1}^N \mu(\boldsymbol{x}_i, \boldsymbol{\theta})$ and, replacing $\boldsymbol{\theta}$ by $\hat{\boldsymbol{\theta}}$, again maximize (11) subject to (12).

Paralleling Chen and Sitter (1999), the Lagrange multiplier method can be used to show that for any finite population parameters that can be written as a functional, $T(F_N)$, the resulting EL is $\hat{T}_n = T(\hat{F}_n)$, where $\hat{F}_n = \sum_{i\in s} \hat{p}_i \delta_{y_i}$, δ_{y_i} is the point measure at y_i, the $\hat{p}_i = w_i/[1 + \lambda u_i]$ for $i \in s$, and the scalar Lagrange multiplier, λ, is the solution to

$$\sum_{i\in s} \frac{w_i u_i}{1 + \lambda u_i} = 0, \qquad (13)$$

where $w_i = d_i/\sum_{i\in s} d_i$. For instance, the EL for $\bar{Y}$ would be $\hat{\bar{Y}}_{EL} = \sum_{i\in s} \hat{p}_i y_i$. Note that no auxiliary information translates into $u_i = 0$ and the resulting EL of $\bar{Y}$ is $\hat{\bar{Y}} = \sum_{i\in s} d_i y_i / \sum_{i\in s} d_i$. One advantage to this approach is that the resulting weights are positive, which may not be true for the other two methods.

A theorem analogous to Theorem 1 of Chen and Sitter (1999) can then be proved.

Theorem 2. *Under the conditions i) - iii) of Theorem 1 and iv) - vi) given below, $\hat{\bar{Y}}_{EL}$, the pseudo-empirical maximum likelihood estimator of $\bar{Y}$ obtained by calibrating on fitted values under model (4), is asymptotically equivalent to the model-calibration estimator $\hat{\bar{Y}}_{MC}$.*

Let $u_i = \mu(\boldsymbol{x}_i, \boldsymbol{\theta}_N) - N^{-1}\sum_{i=1}^N \mu(\boldsymbol{x}_i, \boldsymbol{\theta}_N)$, $h_i = h(\boldsymbol{x}_i, \boldsymbol{\theta}_N)$, where $h(\boldsymbol{x}_i, \boldsymbol{\theta}_N)$ is defined in condition iii) of Theorem 1. The conditions needed are:

iv) $u^* = \max_{i\in s} |u_i| = o_p(n^{1/2})$;

v) $\sum_{i\in s} d_i u_i / \sum_{i\in s} d_i u_i^2 = O_p(n^{-1/2})$;

vi) $h^* = \max_{i\in s} |h_i| = o_p(n)$.

Proof: See Wu (1999).

3.4 Generalized Difference Estimator

The well-known generalized regression estimator (GR) (Cassel, Särndal and Wretman, 1976; Särndal, 1980) can be motivated as a model-assisted generalized difference estimator (GD). Suppose we assume a linear model as in (5) with $\mu_i = \mu(\boldsymbol{x}_i, \boldsymbol{\theta})$ given in (10). The GR can then be written (Särndal, 1980)

as

$$\hat{\bar{Y}}_{GD} = N^{-1}\{\sum_{i\in s} d_i y_i - \sum_{i\in s} d_i \mu(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}) + \sum_{i=1}^{N} \mu(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}})\}. \tag{14}$$

This estimator is obtained by choosing $a_i = \mu(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}})$ in the usual design-based difference estimator $\hat{\bar{Y}}_{DIF} = N^{-1}\{\sum_{i\in s} d_i y_i - \sum_{i\in s} d_i a_i + \sum_{i=1}^{N} a_i\}$. The choice of $a_i = \mu(\boldsymbol{x}_i, \boldsymbol{\theta})$ in $\hat{\bar{Y}}_{DIF}$ is optimal in that it minimizes $E_p\{V_\xi(\hat{\bar{Y}}_{DIF} - \bar{Y})\}$ if $\pi_i \propto v_i$ (Cassel, Särndal and Wretman, 1976). It is fairly clear that, if we put in the form of $\mu(\boldsymbol{x}, \hat{\boldsymbol{\theta}})$ implied by (10), (14) will depend on the x-values only through $\hat{\boldsymbol{X}}_{HT}$ and $\bar{\boldsymbol{X}}$. It is also clear that the motivation generalizes naturally to more complex models by allowing $\mu(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}})$ to come from the development in Section 3.1. However, in this case the resulting estimator depends upon the $\boldsymbol{x}$-values in a more complicated way and in particular complete auxiliary information is necessary to apply the method.

This generalization of $\hat{\bar{Y}}_{GR}$ to $\hat{\bar{Y}}_{GD}$ using model (4) shares many of the nice properties of $\hat{\bar{Y}}_{MC}$. In particular: (i) Theorem 1 can be restated for $\hat{\bar{Y}}_{GD}$ with similar proof; (ii) if $y_i = \mu_i$, $\hat{\bar{Y}}_{GD} = \bar{Y}$; and (iii) under a linear model it reduces to the usual GR.

It is interesting to note that $\hat{\bar{Y}}_{MC}$ and $\hat{\bar{Y}}_{GD}$ are not the same under the general model, as they are under a linear model. The behavior of $\hat{\bar{Y}}_{GD}$ is associated with the "goodness" of approximation $y_i \doteq \mu(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}})$. It depends largely on modeling variation. If $y_i \doteq \mu(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}})$, then $V_p(\hat{\bar{Y}}_{GD}) \doteq 0$. On the other hand, if the relationship between y and $\boldsymbol{x}$ is not strong enough, we might have $V_p(\hat{\bar{Y}}_{GD}) \geq V_p(\hat{\bar{Y}}_{HT})$, no gain by using $\hat{\bar{Y}}_{GD}$. The model-calibration estimator $\hat{\bar{Y}}_{MC}$, on the other hand, uses $\mu(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}})$'s as a tool of calibration while keeping as close to $\hat{\bar{Y}}_{HT}$ as possible. It is arguable that $\hat{\bar{Y}}_{MC}$ will perform much better. $\hat{\bar{Y}}_{MC}$ can be viewed as a regression estimator based on an artificial model $y_i = b_0 + b_1\mu(x_i, \theta) + q_i^{-1/2}\varepsilon_i$. Even in the case of model misspecification, this artificial simple linear regression model might still fit reasonably well and the gain by using a regression estimator is still available.

4 A Simulation

We conducted a limited simulation study to investigate the finite sample performance of the estimators of $\bar{Y}$ proposed in Sections 3.2 through 3.4. A finite population consisting of $N = 2,000$ units was generated as an *iid* sample from $\log(y) = \theta_0 + \theta_1 x + \varepsilon$, where $x \sim Gamma(1,1)$ and $\varepsilon \sim N(0, \sigma^2)$. We chose $\theta_0 = \theta_1 = 1$. Four different finite populations were used by choosing different values of σ^2 such that the correlation coefficient between $\log(y)$ and x are 0.9, 0.8, 0.7 and 0.6, respectively.

For each fixed finite population, a simple random sample of size $n = 100$ was taken and a log-linear model

$$\log(\mu_i) = \alpha + \beta x_i\,, \quad V(\mu) = \mu^2$$

was fit using pseudo maximum quasi-likelihood estimation. Estimators $\hat{\bar{Y}}_{MC}$, $\hat{\bar{Y}}^*_{MC}$, $\hat{\bar{Y}}_{EL}$ and $\hat{\bar{Y}}_{GD}$ were computed using the sample data and all the fitted values. We also included the GR based on a linear model in the simulation to compare to a routine application without modeling. All estimators were compared to the baseline estimator, $\hat{\bar{Y}}_{HT}$. The process was repeated $B = 50,000$ times.

The performance of the various estimators was measured by the simulated Relative Bias (RB, in percentage) and Relative Efficiency (RE), defined by

$$\begin{aligned} RB &= 100 \times B^{-1}\sum_{i=1}^{B}(\hat{\bar{Y}} - \bar{Y})/\bar{Y}\,, \\ RE &= \sqrt{MSE}/\sqrt{MSE_{HT}}\,, \end{aligned} \tag{15}$$

where $MSE = B^{-1}\sum_{i=1}^{B}(\hat{\bar{Y}} - \bar{Y})^2$ and MSE_{HT} is the MSE of $\hat{\bar{Y}}_{HT}$.

Table 1 reports RB and RE for the estimators included in the simulation. Several interesting points are highlighted here: (1) the RB are all within a reasonable range, with the GR having the largest at 5%; (2) $\hat{\bar{Y}}_{MC}$, $\hat{\bar{Y}}^*_{MC}$ and $\hat{\bar{Y}}_{EL}$ perform similarly and better in all cases; (3) $\hat{\bar{Y}}^*_{MC}$ never outperforms $\hat{\bar{Y}}_{MC}$, and $\hat{\bar{Y}}_{EL}$ never outperforms $\hat{\bar{Y}}_{MC}$ or $\hat{\bar{Y}}^*_{MC}$. The reason for the latter may represent the price to be paid to achieve the positive weights; (4) $\hat{\bar{Y}}_{GD}$ performs well when the relationship between y and x is strong (populations 1 and 2), but can be worse than $\hat{\bar{Y}}_{HT}$, which does not even use the auxiliary information, when the relationship is weak (population 4); (5) the gain from using the GR, which ignores the curved relationship between y and x, is always marginal.

5 Concluding Remarks

We have proposed a model-calibration approach to the use of complete auxiliary information in complex surveys to estimate totals and means. The idea

Table 1: Relative Percentage Bias ($RB\%$) and Relative Efficiency (RE)

Population	ρ	$\hat{\bar{Y}}_{HT}$	$\hat{\bar{Y}}_{MC}$	$\hat{\bar{Y}}^{*}_{MC}$	$\hat{\bar{Y}}_{EL}$	$\hat{\bar{Y}}_{GD}$	$\hat{\bar{Y}}_{GR}$
Percentage Relative Bias ($RB\%$)							
1	.9	.15	−1.23	−1.21	−1.12	−.22	−4.97
2	.8	.14	−2.10	−2.14	−2.09	.18	−5.10
3	.7	.07	−2.78	−2.95	−3.05	1.29	−5.27
4	.6	−.33	−2.87	−3.32	−3.60	5.13	−5.71
Relative Efficiency to $\hat{\bar{Y}}_{HT}$ (RE)							
1	.9	1.00	.30	.30	.32	.42	.84
2	.8	1.00	.40	.41	.44	.58	.87
3	.7	1.00	.49	.49	.54	.71	.89
4	.6	1.00	.61	.61	.89	1.13	.90

involves fitting a general working model and then calibrating on the resulting fitted values as opposed to on the auxiliary variables themselves.

We can summarize the innovation in this work through the following points: 1) The relationship between an assumed model and the use of complete auxiliary information is highlighted by noting that, in the case of a linear model it is only necessary to know the mean of the auxiliary variables for the entire finite population to construct efficient estimators of $\bar{Y}$. Therefore, making complete use of auxiliary information requires more complex modeling; 2) The most obvious direction for extending to nonlinear models is through the generalized difference estimator. However, as we demonstrate in the simulation, unless the relationship between y and $\boldsymbol{x}$ is very strong, this approach can do quite poorly and in fact can perform worse than ignoring the auxiliary information all together; and 3) it is not obvious that one can/should avoid calibrating on the vector of auxiliary variables directly. We argue and demonstrate that a simple and powerful way to do this is to calibrate on the fitted values either directly or using a pseudo-empirical likelihood approach.

The use of model-calibration approach for estimating the finite population distribution function and quantiles was discussed by Wu (1999).

References

Basu, D. (1971), *Foundations of Statistical Inference, A Symposium*, eds. V.P Godambe and D.A Sprott, Holt Rinehart and Winston of Canada, Limited, Toronto.

Binder, D.A. (1983), "On the Variances of Asymptotically Normal Estimators from Complex Surveys," *International Statistical Review*, 51, 279-292.

Cassel, C.M., Särndal, C.E., and Wretman, J.H. (1976), "Some Results on Generalized Difference Estimation and Generalized Regression Estimation for Finite Populations," *Biometrika*, 63, 615-620.

Chen, J., and Qin, J. (1993), "Empirical Likelihood Estimation for Finite Populations and the Effective Usage of Auxiliary Information," *Biometrika*, 80, 107-116.

Chen, J., and Sitter, R.R. (1999), "A Pseudo Empirical Likelihood Approach to the Effective Use of Auxiliary Information in Complex Surveys," *Statistica Sinica*, 9, 385-406.

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

Godambe, V.P., and Thompson, M.E. (1986), "Parameters of Superpopulation and Survey Population: Their Relationships and Estimation," *International Statistical Review*, 54, 127-138.

Kovar, J.G., Rao, J.N.K., and Wu, C.F.J. (1988), "Bootstrap and Other Methods to Measure Errors in Survey Estimates," *The Canadian Journal of Statistics*, 16, Supplement, 25-45.

Särndal, C.E. (1980), "On π-inverse Weighting versus Best Linear Unbiased Weighting in Probability Sampling", *Biometrika*, 67, 639-650.

Wu, C. (1999), "The Effective Use of Complete Auxiliary Information From Survey Data", Ph.D. dissertation, Simon Fraser University, Burnaby, BC, Canada.

Zhong, C.X.B., and Rao, J.N.K. (1996), "Empirical Likelihood Inference Under Stratified Random Sampling Using Auxiliary Information" Proceedings of the Section on Survey Research Methods, ASA, Chicago, Illinois, August, 1996. 798-803.