

EFFICIENT ESTIMATION OF QUADRATIC FINITE POPULATION FUNCTIONS

Randy R. Sitter and Changbao Wu¹

ABSTRACT

By viewing quadratic and other second-order finite population functions as totals or means over a derived synthetic finite population, we show that the recently proposed model-calibration and pseudo empirical likelihood methods for effective use of auxiliary information from survey data can be readily extended to obtain efficient estimators of quadratic and other second-order finite population functions. In particular, estimation of a finite population variance, a covariance, or variance of a linear estimator can be greatly improved when auxiliary information is available.

KEY WORDS: Generalized regression estimator; Model-assisted approach; Model-calibration; Pseudo empirical likelihood.

RÉSUMÉ

En regardant les fonctions quadratiques de populations finies ou de second ordre comme totaux ou moyennes au-dessus d'une population finie synthétique dérivée, on montre que le modèle calibré récemment proposé et les pseudo-fonctions de vraisemblance empiriques, pour un usage efficace de l'information auxiliaire provenant des données échantillonales peut être appliqué pour obtenir des estimateurs efficaces des fonctions quadratiques de population finie ou de second ordre. En particulier l'estimation, pour une population finie, de la variance, de la covariance, ou de la variance d'un estimateur linéaire peut être considérablement améliorée lorsque l'information auxiliaire est disponible.

MOTS CLES: Approche basée sur un modèle; estimateur de régression généralisé; étalonnage d'un modèle; pseudo-fonction de vraisemblance empirique.

1. INTRODUCTION

The problem of estimating a finite population mean or total in the presence of auxiliary information has been extensively discussed in survey sampling. While a purely model-based prediction approach has been used by some researchers, the model-assisted approach has gained much popularity in recent literature. Several general procedures have been proposed.

Estimation of quadratic or other higher-order finite population functions is also important. For example, efficient estimators for finite population variances, covariances between two response variables, or variances of linear estimators are highly desirable. Shah and Patel (1996) presented several examples to illustrate why the estimation of population variances and covariances might be useful in their own right. However, due to the relative complexity of these functions, it is not obvious that one can obtain more

efficient estimators for these higher-order population quantities when certain auxiliary information is available from survey data.

In this paper, we develop efficient estimators for quadratic and other second-order finite population functions using the recently proposed model-calibration and model-calibrated pseudo empirical likelihood methods (Wu and Sitter, 2001). The most significant feature of our approach is that it effectively uses auxiliary information at the estimation stage under a general sampling design and a general working model, linear or non-linear, single or multiple variables. The approach is model-assisted in that the resulting estimators are asymptotically design-unbiased irrespective of the correctness of the model, but high efficiency will be achieved if the working model is nearly correct.

¹ Changbao Wu, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON N2L 3G1, cbwu@uwaterloo.ca and Randy Sitter, Department of Statistics, Simon Fraser University, Burnaby, British Columbia, V5A 1S6.

2. EFFICIENT ESTIMATION OF QUADRATIC FUNCTIONS

Let $U = \{1, 2, \dots, N\}$ be the set of labels of the finite population. Associate with unit i are values of response variables, y_i , and covariates, x_i , both vector valued. In the most general case, we assume that the values x_1, \dots, x_N are known for the entire finite population (referred to as complete auxiliary information) but y is known only if the i th unit is selected in the sample, s . However, under a linear working model we only need certain second-order summary statistics of x to be known. Throughout, we assume all the first and second order inclusion probabilities π and π_{ij} are strictly positive. Let $d = 1/\pi$ and $d_{ij} = 1/\pi_{ij}$.

A quadratic finite population function can be defined as $Q = \sum_{i=1}^N \sum_{j=i+1}^N a_{ij} y_i y_j$, where a_{ij} are known constants. More generally, let

$$T = \sum_{i=1}^N \sum_{j=i+1}^N \phi(y_i, y_j),$$

where $\phi(y_i, y_j)$ is a symmetric function (a kernel of degree 2 for a U-statistic). The quadratic form Q defined above and the bilinear parameter discussed in Theberge (1999) are both special cases of T . Some practically useful examples include the finite population variance, the covariance between two response variables, and the Yates-Grundy form of the variance of the Horvitz-Thompson estimator. Other quantities of interest, such as the correlation coefficient or regression coefficient are smooth functions of several quadratic or other second order functions.

If we arrange all the pairs (ij) with $i < j$ in a sequence, and denote this by $\alpha = 1, 2, \dots, N^*$, where $N^* = N(N-1)/2$, we may express T as

$$T = \sum_{\alpha=1}^{N^*} t_\alpha,$$

where $t_\alpha = \phi(y_i, y_j)$ for $\alpha = (ij)$. Thus, T can be viewed as a population total defined on a synthetic finite population $U^* = \{1, 2, \dots, N^*\}$ with characteristic of interest, t_α . The corresponding sample of pairs would be $s^* = \{\alpha = (ij) : i < j \text{ and } i, j \in s\}$. Let $n^* = n(n-1)/2$ be the number of pairs in s^* . The

“first order” inclusion probabilities over this synthetic population are $\pi_\alpha^* = \pi_{ij}$ for unit $\alpha = (ij)$. Let $d_\alpha^* = 1/\pi_\alpha^* = 1/\pi_{ij}$. Note that in many applications the diagonal terms $\phi(y_i, y_i) = 0$. If not, those terms can also be included in T which amounts to changing N^* to $N(N+1)/2$, n^* to $n(n+1)/2$, and using $\pi_{ii} = \pi_i$.

2.1 The model-calibration and the generalized difference estimators

Suppose that the relationship between y_i and x_i can be depicted by a semi-parametric model through the first and second-order moments,

$$E_\varepsilon(y_i | x_i) = \mu(x_i, \theta), \quad (2.1)$$

$$V_\varepsilon(y_i | x_i) = v_i^2 \sigma^2, \quad i = 1, 2, \dots, N,$$

where $\theta = (\theta_0, \dots, \theta_p)'$ and σ^2 are unknown superpopulation parameters, $\mu(x, \theta)$ is a known function of x and θ , the v_i is a known function of x_i or $\mu_i = \mu(x_i, \theta)$, and E_ε and V_ε denote the expectation and variance with respect to the superpopulation model. We also assume that $(y_1, x_1), \dots, (y_N, x_N)$ are mutually independent. We restrict to scalar y_i at this point for the ease of presentation.

Let θ_N be an estimate of θ based on the entire finite population, and let $y_i^* = \mu(x_i, \theta_N)$. If the working model (2.1) is appropriate, y_i^* should be an “ideal” choice for approximating y_i , i.e., y_i^* should have higher linear correlation with y_i than other ad hoc choices if the model is adequate and the auxiliary information is really informative.

A design-based estimator \hat{T} for T can be obtained from the sample data (Wu and Sitter, 2001). Let $\hat{y}_i = \mu(x_i, \hat{\theta})$, $i = 1, 2, \dots, N$ be the fitted values from model (2.1). The model-calibration method of Wu and Sitter (2001) can be easily extended to the present context for the estimation of T by using fitted values $\hat{t}_\alpha = \phi(\hat{y}_i, \hat{y}_j)$ for t_α and treating $d_\alpha^* = 1/\pi_{ij}$ as the basic design weights, where $\alpha = (ij)$. Recall that $d_{ij} = 1/\pi_{ij}$ and use the original pair index (ij) , the

model-calibration estimator of the quadratic function T is then defined as $\hat{T}_{MC} = \sum \sum_{(ij) \in s} \omega_{ij} \phi(y_i, y_j)$, where the ω_{ij} 's minimize an average distance measure between ω_{ij} and d_{ij} subject to

$$\begin{aligned} \sum_{i \in s} \sum_{j > i} \omega_{ij} &= N^* \\ \sum_{i \in s} \sum_{j > i} \omega_{ij} \phi(\hat{y}_i, \hat{y}_j) &= \sum_{i=1}^N \sum_{j=i+1}^N \phi(\hat{y}_i, \hat{y}_j). \end{aligned} \quad (2.2)$$

Under the simple chi-square distance measure

$$\Phi_s = \sum_{i \in s} \sum_{j > i} (\omega_{ij} - d_{ij})^2 / (d_{ij} q_{ij}),$$

where the q_{ij} 's are known positive weights unrelated to d_{ij} , the resulting estimator is given by

$$\begin{aligned} \hat{T}_{MC} &= \sum_{i \in s} \sum_{j > i} d_{ij} \phi(y_i, y_j) \\ &+ \left\{ \sum_{i=1}^N \sum_{j=i+1}^N \phi(\hat{y}_i, \hat{y}_j) - \sum_{i \in s} \sum_{j > i} d_{ij} \phi(\hat{y}_i, \hat{y}_j) \right\} \hat{B}, \end{aligned} \quad (2.3)$$

where $\hat{B} = C(u, v) / C(u, u)$, $C(u, v) = \sum \sum_{(ij) \in s} d_{ij} q_{ij} (u_{ij} - \bar{u})(v_{ij} - \bar{v})$, $u_{ij} = \phi(\hat{y}_i, \hat{y}_j)$, $v_{ij} = \phi(y_i, y_j)$, $\bar{u} = \sum \sum_{(ij) \in s} d_{ij} q_{ij} u_{ij} / \sum \sum_{(ij) \in s} d_{ij} q_{ij}$, \bar{v} and $C(u, u)$ are similarly defined. Note that the first term on the right hand side of (2.3) is the usual Horvitz-Thompson estimator, \hat{T}_{HT} . For simplicity of presentation, we consider single y variable with model (2.1) in the following theorem but the results hold for the general case of vector response variables. Under proper asymptotic settings, assuming $T = O_p(N^*)$, we have

Theorem 1, 1) Under certain regularity conditions, the MC estimator $\hat{T}_{MC} = \hat{T}_{HT} + O_p(N^* / \sqrt{n^*})$ and is therefore an asymptotically design unbiased estimator of T .

2) The asymptotic design-based variance of \hat{T}_{MC} is given by $Var(\hat{T}_{MC}) =$

$$\frac{1}{2} \sum_{i=1}^N \sum_{j=i+1}^N \sum_{l=1}^N \sum_{m=l+1}^N (\pi_{ij} \pi_{lm} - \pi_{ijlm}) \left(\frac{E_{ij}}{\pi_{ij}} - \frac{E_{lm}}{\pi_{lm}} \right)^2,$$

where π_{ijlm} are the fourth-order inclusion probabilities, $E_{ij} = v_{ij} - u_{ij} B_N$ and $B_N = C_N(u, v) / C_N(u, v)$,

$C_N(u, v) = \sum_{i=1}^N \sum_{j=i+1}^N d_{ij} q_{ij} (u_{ij} - \bar{u}_N)(v_{ij} - \bar{v}_N)$, $u_{ij} = \phi(y_i^*, y_j^*)$, $v_{ij} = \phi(y_i, y_j)$, $y_i^* = \mu(x_i, \theta_N)$, $\bar{u} = \sum_{i=1}^N \sum_{j=i+1}^N u_{ij} / N^*$, $C_N(u, u)$ and \bar{v} are similarly defined.

The required regularity conditions and a proof of the Theorem can be found in Sitter and Wu (2000).

The relationship between $Var(\hat{T}_{MC})$ and $Var(\hat{T}_{HT})$ is similar to that of $Var(\hat{Y}_{GR})$ and $Var(\hat{Y}_{HT})$, where \hat{Y}_{GR} is the GREG estimator for the population total. The E_{ij} 's are the fitted residuals of the "response variable" $t_\alpha = v_{ij}$ over "covariate" u_{ij} . The fact that the E_{ij} 's are less variable than the v_{ij} 's implies that $Var(\hat{T}_{MC})$ will be smaller than $Var(\hat{T}_{HT})$ for most commonly used sampling designs. The reduction of the variance depends on the correlation coefficient $\rho_{v,u}$ between v_{ij} and u_{ij} . The two extreme cases are (a) $|\rho_{v,u}| = 1$ where $E_{ij} = 0$ and $Var(\hat{T}_{MC}) = 0$; and (b) $\rho_{v,u} = 0$ where $E_{ij} = v_{ij} - \bar{v}$ and $Var(\hat{T}_{MC}) = Var(\hat{T}_{HT})$. This relationship can be seen more clearly under simple random sampling where it is easy to show that

$$Var(\hat{T}_{MC}) = Var(\hat{T}_{HT})(1 - \rho_{v,u}^2) + O(N^* / n^*).$$

Note that $Var(\hat{T}_{HT}) = O(N^{*2} / n^*)$, the MC estimator of T will perform at least as good as HT estimator, with the amount of reduction of variance depending on $\rho_{v,u}$.

If x provides relevant information in explaining y through a model like (2.1), i.e., the correlation between y_i and $y_i^* = \mu(x_i, \theta_N)$ is high, and consequently a high correlation between v_{ij} and u_{ij} , the gain from using \hat{T}_{MC} over \hat{T}_{HT} can be substantial. Moreover, the construction of \hat{T}_{MC} requires no extra step in the modelling stage. The same fitted values \hat{y}_i are used for the estimation of any quadratic and other second-order population functions. Indeed, the structure of (2.3) is identical to the generalized regression

estimator applied to a single response variable $t_\alpha = v_{ij}$ and a single auxiliary variable $t_\alpha^* = \phi(y_i^*, y_j^*)$.

Computation for \hat{T}_{MC} requires no extra matrix manipulation after the initial modelling and therefore is extremely simple.

It is interesting to notice that if we let $\hat{B} = 1$ in (2.3), the resulting estimator can be viewed as a generalized difference (GD) estimator as discussed above. We denote this estimator by \hat{T}_{GD} .

The \hat{T}_{GD} is computationally simpler than \hat{T} and can perform very well under correctly specified working model where the correlation between t_α and t_α^* is high. It may also perform poorly under misspecified superpopulation model, where the fitted values are off the target. The model-calibration estimator, on the other hand, is more robust.

2.2 The pseudo empirical maximum likelihood estimator

We define the model-calibrated pseudo-empirical maximum likelihood estimator of T as

$$\hat{T}_{EL} = N^* \sum_{(ij) \in S} \hat{p}_{ij} \phi(y_i, y_j)$$

where the \hat{p}_{ij} 's maximize

$$\hat{l}(p) = \sum_{i \in S} \sum_{j > i} d_{ij} \log p_{ij},$$

subject to constraints

$$\sum_{i \in S} \sum_{j > i} p_{ij} = 1 \quad (p_{ij} > 0),$$

$$\sum_{i \in S} \sum_{j > i} p_{ij} \phi(\hat{y}_i, \hat{y}_j) = \frac{1}{N^*} \sum_{i=1}^N \sum_{j=i+1}^N \phi(\hat{y}_i, \hat{y}_j). \quad (2.4)$$

Theorem 2. Under proper regularity conditions, the pseudo-empirical maximum likelihood estimator of T , \hat{T}_{EL} , is asymptotically equivalent to the model-calibration estimator \hat{T}_{MC} under the uniform weight $q_{ij} = 1$, i.e. $\hat{T}_{EL} = \hat{T}_{MC} + o_p(N^* / \sqrt{n^*})$. Thus \hat{T}_{EL} is also an asymptotically design-unbiased estimator of T with the same asymptotic variance as \hat{T}_{MC} .

Proof of the Theorem can be found in Sitter and Wu (2000).

One of the advantages of \hat{T} , in addition to its likelihood-based motivation, is that the weights, \hat{p}_{ij} , are always positive which may not be true for the weights of the model-calibration estimator, w_{ij} . This property might be very attractive when all the terms in T are themselves positive and therefore positive weights will ensure positive estimation. A simple and stable algorithm for computing \hat{T} can be found in Sitter and Wu (2000).

3. ESTIMATING VARIANCES AND COVARIANCES UNDER A LINEAR MODEL

In this section we consider the estimation of

$$S_y^2 = (N-1)^{-1} \sum_{i=1}^N (y_i - \bar{Y})^2$$

and

$$C_{yz} = (N-1)^{-1} \sum_{i=1}^N (y_i - \bar{Y})(z_i - \bar{Z})$$

using the proposed estimation strategy. We consider linear regression working models for both y and z , i.e., $E_\epsilon(y_i) = \chi_i' \beta$ and $E_\epsilon(z_i) = \chi_i' \gamma$. The fitted values for y_i and z_i would be $\hat{y}_i = x_i' \hat{\beta}$ and $\hat{z}_i = x_i' \hat{\gamma}$, respectively, where $\hat{\beta}$ and $\hat{\gamma}$ are the design-based estimator for the regression coefficients β and γ . It is then straightforward to show that the model-calibration estimators for S_y^2 and C_{yz} are given by

$$\hat{S}_{MC}^2 = \hat{S}_{HT}^2 + \hat{\beta}' (S_x^2 - s_x^2) \hat{\beta} \hat{B}_1,$$

$$\hat{C}_{MC}^2 = \hat{C}_{HT}^2 + \hat{\beta}' (S_x^2 - s_x^2) \hat{\gamma} \hat{B}_2,$$

where

$$\hat{S}_{HT}^2 = \frac{1}{N(N-1)} \sum_{i \in S} \sum_{j > i} d_{ij} (y_i - y_j)^2,$$

$$\hat{C}_{HT} = \frac{1}{N(N-1)} \sum_{i \in S} \sum_{j > i} d_{ij} (y_i - y_j)(z_i - z_j),$$

$$S_x^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{X})',$$

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

and \hat{B}_1, \hat{B}_2 are similarly defined as \hat{B} in (2.3) with $u_{ij} = \hat{\beta}'(x_i - x_j)(x_i - x_j)'\hat{\beta}$ and $v_{ij} = (y_i - y_j)^2$

for \hat{B}_1 , $u_{ij} = \hat{\beta}'(x_i - x_j)(x_i - x_j)'\hat{\gamma}$ and $v_{ij} = (y_i - y_j)(z_i - z_j)$ for \hat{B}_2 .

The generalized difference estimators \hat{S}_{GD}^2 and \hat{C}_{GD}^2 are obtained by letting $\hat{B}_1 = \hat{B}_2 = 1$, above. The PEML estimators for S_y^2 and C_{yz} are given by

$$\hat{S}_{EL}^2 = \frac{1}{2} \sum_{i \in S} \sum_{j > i} \hat{p}_{ij} (y_i - y_j)^2,$$

$$\hat{C}_{EL} = \frac{1}{2} \sum_{i \in S} \sum_{j > i} \hat{p}_{ij} (y_i - y_j)(z_i - z_j),$$

where $N^* = N(N-1)/2$, $\hat{p}_{ij} = d_{ij}^*/(1 + \lambda b_{ij})$ and λ is the Lagrange multiplier.

One should note that in the case of a linear working model, we see that, despite the motivation of using the predicated values for each unit and thus implicitly desiring complete auxiliary information, one only requires knowledge of certain population quantities of the x variables such as S_x^2 to construct the proposed estimators.

4. SOME EMPIRICAL RESULTS

In this section, a limited simulation study is carried out to investigate the finite sample performance of the proposed variance estimators using the 1996 Statistics Canada's Family Expenditure (FAMEX) Survey data for the province of Ontario, down-loaded from Statistics Canada Databases. The data contains observations on $N = 2396$ sampled households over a variety of variables. For the purpose of illustration, we choose x_1 : number of people in the household and x_2 : total income after taxes as auxiliary variables, y_1 : annual expenditures on clothing and y_2 : total expenditure as the study variables. Since extra information is not available to us, we treat this data set itself as the finite population in the simulation study. The population is split into eight strata using the original design weights.

A scatter plot reveals that a linear working model might be appropriate, but the relationship between the x variables and the y variables is not particularly strong. The finite population correlation coefficients are $\rho_{x_1, y_1} = 0.40$, $\rho_{x_1, y_2} = 0.44$, $\rho_{x_2, y_1} = 0.60$ and $\rho_{x_2, y_2} = 0.87$.

A stratified simple random sample of size $n = 64$, with 8 units from each stratum, was drawn and MC , GD and EL estimators of Section 3 for the population variances $S_{y_1}^2$ and $S_{y_2}^2$ and covariance $S_{12} = Cov(y_1, y_2)$ were computed. The uniform weights $q_{ij} = 1$ were used for the MC estimators. The usual HT estimators were also included for comparison. This process was repeated $B = 1,000$ times. Note that the sampling fraction here is approximately 2.5%.

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

$$RB = 100 \times B^{-1} \sum_{b=1}^B (\hat{S}_b^2 - S^2) / S^2$$

and $RE = MSE_{HT} / MSE$, where \hat{S}_b^2 is the estimate of S^2 computed from the b th simulated sample,

$$MSE = B^{-1} \sum_{b=1}^B (\hat{S}_b^2 - S^2)^2 \text{ and } MSE_{HT} \text{ is the } MSE \text{ of } \hat{S}_{HT}^2.$$

Table 1 reports the RB and RE for estimators included in the simulation. The absolute values of RB's are all within reasonable range, with the largest occurring for the EL estimators at 5.664%. In terms of efficiency, all of MC , GD and EL estimators outperform the conventional HT estimators, with MSE reduced by almost half for estimating the covariance S_{12} . The MC and EL estimators perform better than the GD estimator when estimating $S_{y_1}^2$ and S_{12} but this is reversed when estimating $S_{y_2}^2$. One possible reason

Table 1. Relative Bias(RE) & Relative Efficiency(RE)

Parameter	MC	GD	EL
	Percent	Relative	Bias (RB)
$S_{y_1}^2$	-3.259	-.993	-3.759
$S_{y_2}^2$	4.743	3.743	5.664
S_{12}	.625	1.346	.022
	Relative	Efficiency	to HT (RE)
$S_{y_1}^2$	1.629	1.454	1.648
$S_{y_2}^2$	1.318	1.525	1.289
S_{12}	1.913	1.852	1.956

for this is that the linear correlation between y_2 and (x_1, x_2) is much stronger than that of y_1 ($\rho_{x_2, y_1} = 0.60$, $\rho_{x_2, y_2} = 0.87$). The GD estimator usually performs well in this case.

5. CONCLUDING REMARKS

One should note that the finite population variance can be written as

$$S_y^2 = (N-1)^{-1} [\sum_{i=1}^N y_i^2 - (\sum_{i=1}^N y_i)^2 / N].$$

The two population totals $\sum_{i=1}^N y_i^2$ and $\sum_{i=1}^N y_i$ could then be estimated separately using model-calibration or pseudo empirical maximum likelihood methods. The resulting estimators would involve no second-order inclusion probabilities, however, the non-negativity of the estimators obtained from the two separate pieces would not be guaranteed.

Some attractive features of the proposed methodology are: 1) estimators can be constructed under a general sampling design; 2) the methods can handle scalar or vector valued auxiliary variables, as well as, linear or non-linear working models and are therefore very flexible; 3) the approach is model-assisted in that the resulting estimators are asymptotically design-unbiased regardless of the correctness of the model but have high efficiency if the working-model adequately describes the relationship between the response variables and the covariates; 4) the approaches require no additional step for the modelling, the same fitted values are used for any quadratic or other second order population functions; 5) in the case of linear working models, estimation of population variances and covariances or the variance of a linear estimator requires only the S_x^2 or other second-order summary

statistics of x be known at the population level. This is much like the GREG for the population mean where only \bar{X} need to be known to construct the estimator; 6) after the initial modelling stage (e.g. estimation of β and γ), the construction of proposed estimators involve only scalar manipulations and the resulting estimators have very simple forms; 7) the limited empirical results show that the proposed estimators for population variances and covariances are very efficient compared to the conventional estimators; and 8) the model-calibrated pseudo empirical maximum likelihood estimators ensure nonnegative estimation for certain positive quantities such as population variances.

REFERENCES

- Shah, D.N. and Patel, P.A. (1996), "Asymptotic Properties of a Generalized Regression type Predictor of a Finite Population Variance in Probability Sampling." *Canadian J. Statist.*, **24**, 373-384
- Sitter, R.R. and Wu, C. (2000), "Efficient Estimation of Quadratic Finite Population Functions in the Presence of Auxiliary Information", Working paper 2000-09, department of Statistics and Actuarial Science, University of Waterloo.
- Theberge, A. (1999), "Extensions of Calibration Estimators in Survey Sampling." *J. Amer. Statist. Assoc.*, **94**, 635-644.
- Wu, C. and Sitter, R.R. (2001), "A Model-Calibration Approach to Using Complete Auxiliary Information From Survey Data." *J. Amer. Statist. Assoc.*, **96**, 185-193.