Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Model-assisted calibration estimation using generalized entropy calibration in survey sampling

by Jae Kwang Kim, Yonghyun Kwon, Yumou Qiu and Junyong Park

Release date: June 30, 2025

Statistics Canada   Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at infostats@statcan.gc.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-514-283-9350

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, the Agency has developed standards of service which its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

Survey Methodology, June 2025
Vol. 51, No. 1, pp. 197-215
Statistics Canada, Catalogue No. 12-001-X

Model-assisted calibration estimation using generalized entropy calibration in survey sampling

Jae Kwang Kim, Yonghyun Kwon, Yumou Qiu and Junyong Park[1]

Abstract

We introduce a novel approach to model-assisted calibration estimation in survey sampling using generalized entropy. The method builds upon recent work by Kwon, Kim and Qiu (2024) and extends it to a model-assisted framework. Unlike traditional calibration techniques, this approach employs a generalized entropy function as the objective for optimization and incorporates a debiasing calibration constraint to ensure design consistency. The proposed estimator is shown to be asymptotically equivalent to an augmented generalized regression (GREG) estimator. It allows for unequal model variance, potentially improving efficiency when the sampling design is informative. The paper presents both design-based and model-based justifications for the method, along with asymptotic properties and variance estimation techniques. Computational aspects are discussed, including an unconstrained optimization approach that facilitates implementation, especially for high-dimensional auxiliary variables. The method's performance is evaluated through a simulation study, demonstrating its effectiveness in improving estimation efficiency, particularly when the sampling design is informative.

Key Words: Bias reduction; Calibration weighting; Debiased projection; Informative sampling design; Regression estimation.

1. Introduction

Incorporating auxiliary information to improve the efficiency of design-based estimators is an important practical problem in survey sampling. By properly integrating auxiliary information into the final estimation, the resulting estimator can be more efficient. A popular method of incorporating auxiliary information is calibration weighting. See Fuller (2002), Kim and Park (2010) and Kott (2015) for reviews of literature on calibration weighting.

Although calibration weighting implicitly uses a model, achieving design consistency is crucial, as the underlying regression model may not be entirely accurate. The model-assisted approach is attractive because it uses the model to motivate the estimator while maintaining a design-based mode of inference. For this reason, model-assisted approach is widely accepted in survey sampling (Särndal, Swensson and Wretman, 1992; Breidt and Opsomer, 2017). If the model is used explicitly in the calibration, model calibraion of Wu and Sitter (2001) can be considered.

Many papers on calibration weighting are based on the framework of Deville and Särndal (1992), which minimizes a distance measure between the design weights and the final weights subject to calibration constraints. When the sample size is sufficiently large, the law of large numbers ensures that the calibration constraint is almost satisfied. In this case, the final weights are nearly identical to the original design weights. Consequently, as the sample size increases, the final weights converge to the design weights, justifying design consistency. When the sample size is not sufficiently large, the calibration constraints help reducing

1. Jae Kwang Kim, Department of Statistics, Iowa State University, Ames, Iowa, U.S.A. E-mail: jkim@iastate.edu; Yonghyun Kwon, Department of Mathematics, Korea Military Academy, Seoul, Republic of Korea; Yumou Qiu, School of Mathematical Sciences, Peking University, Beijing, China; Junyong Park, Department of Statistics, Seoul National University, Seoul, Republic of Korea.

the variance of the final estimator, provided that the variables used in the calibration are related to the study variable. Therefore, the calibration weighting using the framework of Deville and Särndal (1992) has a model-assisted justification.

Recently, Kwon et al. (2024) proposed a novel framework of calibration weighting that does not use the traditional distance measure between two weights. Instead, they proposed using a generalized entropy function (Gneiting and Raftery, 2007) as the objective function for optimization and adding a so-called debiasing calibration constraint. The debiasing calibration constraint is key to achieving the design consistency of the final calibration estimator.

In this paper, we introduce the new calibration weighting method of Kwon et al. (2024) within the model-assisted framework. Unlike the generalized regression estimator, the calibration weighting estimator uses regression models implicitly. Nonetheless, the model-assisted framework can be applied to this proposed calibration weighting method. The superpopulation model we consider allows for unequal model variance, which can further improve the efficiency of the resulting calibration estimator. In addition to discussing its design-based asymptotic properties, we also consider finite-sample properties under the superpopulation model and develop a modified variance estimator that improves finite-sample performance. Some computational details are also discussed.

The paper is organized as follows. In Section 2, basic setup and the research problems are introduced. In Section 3, model-assisted calibration is introduce. The proposed method is introduced in Section 4 and its asymptotic properties are presented in the design-based framework. In Section 5, a model-based justification is made. Some computational details are covered in Section 6. Results from a limited simulation study are presented in Section 7. Some concluding remarks are made in Section 8.

2. Basic setup

We consider the situation where certain auxiliary variables are observed throughout the finite population and the study variables are observed only in the sample. We use $\mathbf{x}_i \in \mathbb{R}^p$ and y_i to denote the vector of auxiliary variables and the study variable associated with unit i, respectively. Assume that we have a probability sample $A \subset U$ selected from the finite population and observed y_i in the sample. Let π_i be the first-order inclusion probability of unit i. We assume that π_i are available throughout the finite population. To utilize the auxiliary information $\mathbf{x}_i$ observed throughout the finite population, one can impose the following superpopulation model

$$y_i = \mathbf{x}_i' \boldsymbol{\beta} + e_i, \tag{2.1}$$

where $\mathbf{x}_i$ includes an intercept term, $\boldsymbol{\beta} \in \mathbb{R}^p$ is unknown model parameter and e_i satisfies $E(e_i) = 0,$ $E(e_i \mathbf{x}_i) = 0,$ and $V(e_i \mid \mathbf{x}_i) := v_i,$ where v_i is a known function of $\mathbf{x}_i$. The regression superpopulation model in (2.1) is not necessarily correct but motivates our proposed estimator utilizing the auxiliary variable $\mathbf{x}_i$ observed throughout the finite population.

We are interested in estimating $Y = \sum_{i=1}^{N} y_i$ from the sample. We consider a linear estimator defined as

$$\hat{Y}_\omega = \sum_{i \in A} \omega_i y_i$$

for some $\omega_i > 0$ that does not depend on y-values.We impose that the final weight satisfy the calibration constraint:

$$\sum_{i \in A} \omega_i \mathbf{x}_i = \sum_{i=1}^{N} \mathbf{x}_i. \tag{2.2}$$

Now, under (2.1), note that

$$\hat{Y}_\omega - Y = \left\{ \sum_{i \in A} \omega_i \mathbf{x}_i' \boldsymbol{\beta} - \sum_{i=1}^{N} \mathbf{x}_i' \boldsymbol{\beta} \right\} + \left\{ \sum_{i \in A} \omega_i e_i - \sum_{i=1}^{N} e_i \right\} := C + D. \tag{2.3}$$

The first term, $C,$ can be eliminated if the weights ω_i satisfy the calibration constraint (2.2). Consequently, we have only to minimize the model variance of the term $D.$ Note that

$$E_\zeta (D^2 \mid A) = \sum_{i \in A} \omega_i^2 v_i - 2 \sum_{i \in A} \omega_i v_i + \sum_{i=1}^{N} v_i,$$

where subscript ζ represents the superpopulation model in (2.1). Thus, the conditional expectation in $E_\zeta (\cdot \mid A)$ represent the model expectation conditional on the realized sample $A.$ If the condition $v_i = \lambda' \mathbf{x}_i$ holds for some $\lambda,$ then the calibration constraint in (2.2) implies that

$$\sum_{i \in A} \omega_i v_i = \sum_{i=1}^{N} v_i \tag{2.4}$$

and we obtain

$$E_\zeta (D^2 \mid A) = \sum_{i \in A} \omega_i^2 v_i - \sum_{i=1}^{N} v_i.$$

Therefore, minimizing the model variance of D under constraint (2.2) is equivalent to minimizing

$$Q(w) = \sum_{i \in A} \omega_i^2 v_i \tag{2.5}$$

subject to the same constraint. The optimal calibration estimator that minimizes (2.5) subject to (2.2) is then given by

$$\hat{Y}_{\text{proj}} = \sum_{i \in A} \hat{\omega}_i y_i = \sum_{i=1}^{N} \mathbf{x}_i' \hat{\boldsymbol{\beta}}_v \tag{2.6}$$

where

$$\hat{\omega}_i = \sum_{i=1}^{N} \mathbf{x}_i' \left(\sum_{i \in A} v_i^{-1} \mathbf{x}_i \mathbf{x}_i' \right)^{-1} v_i^{-1} \mathbf{x}_i$$

and

$$\hat{\boldsymbol{\beta}}_v = \left(\sum_{i \in A} v_i^{-1} \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \sum_{i \in A} v_i^{-1} \mathbf{x}_i y_i.$$

This formulation reveals that the final calibration estimator can be interpreted as a projection estimator under model (2.1). In other words, the calibration estimator using constraint (2.2) uses a regression super-population model in (2.1) implicitly.

To achieve the design consistency, note that the approximate design-bias of the projection estimator is

$$\text{Bias}(\hat{Y}_{\text{proj}} \mid \mathcal{F}_N) = E\left(\hat{Y}_{\text{proj}} \mid \mathcal{F}_N\right) - Y \doteq \sum_{i=1}^{N} \left\{ \mathbf{x}_i' \boldsymbol{\beta}_v - y_i \right\}$$

where

$$\boldsymbol{\beta}_v = \left(\sum_{i=1}^{N} \pi_i v_i^{-1} \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \sum_{i=1}^{N} \pi_i v_i^{-1} \mathbf{x}_i y_i$$

is the probability limit of $\hat{\boldsymbol{\beta}}_v$ in (2.6) and $\mathcal{F}_N = \{(\mathbf{x}_i, y_i);\ i = 1, \ldots, N\}$. The conditional expectation conditional on $\mathcal{F}_N$ represents the expectation with respect to the sampling mechanism. Thus, we can estimate the bias by

$$\widehat{\text{Bias}}(\hat{Y}_{\text{proj}}) = \sum_{i \in A} \frac{1}{\pi_i} \left(\mathbf{x}_i' \hat{\boldsymbol{\beta}}_v - y_i \right)$$

and obtain the debiased projection estimator

$$\begin{aligned} \hat{Y}_{\text{d, proj}} &= \hat{Y}_{\text{proj}} - \widehat{\text{Bias}}(\hat{Y}_{\text{proj}}) \\ &= \sum_{i=1}^{N} \mathbf{x}_i' \hat{\boldsymbol{\beta}}_v + \sum_{i \in A} \frac{1}{\pi_i} \left(y_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}}_v \right). \end{aligned} \tag{2.7}$$

The debiased projection estimator can be treated as a variant of the generalized regression (GREG) estimator (Deville and Särndal, 1992). The choice of $\hat{\boldsymbol{\beta}}$ in the debiased projection estimator is not critical as the effect of $\hat{\boldsymbol{\beta}}$ in the debiased projection estimator asymptotically negligible. While the projection term is efficient under the superpopulation model, the bias correction term is estimated using the design weights, which does not necessarily lead to efficient estimation. We consider an alternative approach to debiased projection estimator using model-assisted calibration in the next section.

3. Model-assisted calibration

In the previous section, the model-based projection estimator is modified to achieve the design consistency by substracting a design-based estimator of the bias term. Another way of achieving the design consistency in the projection estimator is to augment the covariate in the projection estimator (Särndal and

Wright, 1984). Specifically, we can augment $\mathbf{x}_i$ to include $\pi_i^{-1} v_i$. In other words, by defining $\mathbf{z}_i' = (\mathbf{x}_i', \pi_i^{-1} v_i)$ and calculating $\hat{\boldsymbol{\gamma}} = \left(\sum_{i \in A} v_i^{-1} \mathbf{z}_i \mathbf{z}_i'\right)^{-1} \sum_{i \in A} v_i^{-1} \mathbf{z}_i y_i$, the property of the residuals ensures that $\hat{y}_i = \mathbf{z}_i' \hat{\boldsymbol{\gamma}}$ fulfills

$$\sum_{i \in A} \left(y_i - \hat{y}_i\right) \mathbf{z}_i v_i^{-1} = \mathbf{0},$$

thereby implying

$$\sum_{i \in A} \frac{1}{\pi_i} \left(y_i - \hat{y}_i\right) = 0. \tag{3.1}$$

Condition (3.1) makes the bias correction term identically equal to zero, which implies that the projection estimator $\hat{Y}_{\text{proj}} = \sum_{i=1}^{N} \hat{y}_i$ is already design consistent. In this sense, the condition (3.1) can be called the *internal bias calibration* (IBC) condition, which is termed by Firth and Bennett (1998).

Using the algebraic equivalence between the projection estimator and the calibration estimator, we can express the projection estimator using the augmented covariate $\mathbf{z}_i$ as a calibration estimator. That is, to minimize the model variance $\hat{\theta}_\omega$ while also adhering to the design consistency condition, our objective narrows down to minimizing

$$\sum_{i \in A} \omega_i^2 v_i \tag{3.2}$$

subject to

$$\sum_{i \in A} \omega_i \mathbf{z}_i = \sum_{i=1}^{N} \mathbf{z}_i \tag{3.3}$$

where $\mathbf{z}_i' = (\mathbf{x}_i', \pi_i^{-1} v_i)$. Including $\pi_i^{-1} v_i$ into calibration corresponds to the IBC condition in the projection estimator.

Let $\hat{\omega}_i$ be the solution to the optimization problem described above. Using the Lagrangian multiplier method, it can be shown that the resulting estimator can be written as

$$\hat{Y}_{\text{cal}} = \sum_{i \in A} \hat{\omega}_i y_i = \sum_{i=1}^{N} \mathbf{z}_i' \hat{\boldsymbol{\gamma}}, \tag{3.4}$$

where

$$\hat{\omega}_i = \left(\sum_{i=1}^{N} \mathbf{z}_i\right)' \left(\sum_{i \in A} v_i^{-1} \mathbf{z}_i \mathbf{z}_i'\right)^{-1} \mathbf{z}_i / v_i$$

and

$$\hat{\boldsymbol{\gamma}} = \left(\sum_{i \in A} v_i^{-1} \mathbf{z}_i \mathbf{z}_i'\right)^{-1} \sum_{i \in A} v_i^{-1} \mathbf{z}_i y_i.$$

After some linear algebra, we can express (3.4) as

$$\hat{Y}_{\text{cal}} = \sum_{i \in U} \mathbf{x}_i' \hat{\boldsymbol{\gamma}}_1 + \left(\frac{\sum_{i \in U} z_{2i}}{\sum_{i \in A} \pi_i^{-1} z_{2i}} \right) \sum_{i \in A} \frac{1}{\pi_i} \left(y_i - \mathbf{x}_i' \hat{\boldsymbol{\gamma}}_1 \right), \tag{3.5}$$

where $z_{2i} = \pi_i^{-1} v_i$, $\hat{\boldsymbol{\gamma}}_1 = \left(\sum_{i \in A} v_i^{-1} \tilde{\mathbf{z}}_{1i} \tilde{\mathbf{z}}_{1i}' \right)^{-1} \sum_{i \in A} v_i^{-1} \tilde{\mathbf{z}}_{1i} y_i$, and $\tilde{\mathbf{z}}_{1i} = \mathbf{x}_i - z_{2i} \left(\sum_{i \in A} v_i^{-1} z_{2i}^2 \right)^{-1} \left(\sum_{i \in A} v_i^{-1} z_{2i} \mathbf{x}_i \right)$. Comparing (3.5) with (2.7), the bias-correction term is further multiplied to a ratio adjustment term $\hat{R}_{z2} = \left(\sum_{i \in A} \pi_i^{-1} z_{2i} \right)^{-1} \sum_{i \in U} z_{2i}$. The ratio adjustment term can improve the efficiency if the correlation between $\hat{e}_i = y_i - \mathbf{x}_i' \hat{\boldsymbol{\gamma}}_1$ and $z_{2i} = \pi_i^{-1} v_i$ is high (Fuller, 2009). If the regression model in (2.1) is correct and the sampling mechanism is non-informative in the sense of Pfeffermann and Sverchkov (2009), then the correlation is nearly zero and there is no efficiency gain of using (3.5) over the GREG estimator in (2.7). On the other hand, if the sampling design is informative under the superpopulation model in (2.1), then $z_{2i} = \pi_i^{-1} v_i$ may contain extra information and the ratio-adjustment term in (3.5) can further improve the efficiency of the resulting estimator.

Let $\boldsymbol{\gamma}^*$ be the probability limit of $\hat{\boldsymbol{\gamma}}$. Using the standard argument, we can obtain

$$N^{-1} \hat{Y}_{\text{cal}} = N^{-1} \hat{Y}_{\text{diff}} + o_p(n^{-1/2}),$$

where

$$\hat{Y}_{\text{diff}} = \sum_{i \in U} \mathbf{z}_i' \boldsymbol{\gamma}^* + \sum_{i \in A} \frac{1}{\pi_i} \left(y_i - \mathbf{z}_i' \boldsymbol{\gamma}^* \right).$$

Now, note that

$$E\left\{ \left(\hat{Y}_{\text{diff}} - Y \right)^2 \right\} = \sum_{i=1}^{N} \left(\frac{1}{\pi_i} - 1 \right) v_i^*$$

where $v_i^* = E_\zeta \left\{ (y_i - \mathbf{z}_i' \boldsymbol{\gamma}^*)^2 \mid \mathbf{z}_i \right\}$. Thus, if the regression model specified in (2.1) is accurate, then $v_i^* = v_i$. However, if the regression model does not accurately represent the underlying data structure, it is possible to observe that $v_i^* < v_i$. This discrepancy arises because the additional covariate $\pi_i^{-1} v_i$ improves the prediction of y_i, thus contributing to a more precise estimation.

4. Proposed method

Instead of the squared error loss in (2.5), we now consider maximizing the generalized entropy that does not depend on the design weights, which was proposed by Kwon et al. (2024). Let $G: \mathcal{V} \to \mathbb{R}$ be a pre-specified function that is strictly convex and twice-continuously differentiable. The domain of G is an open interval $\mathcal{V} = (\nu_1, \nu_2)$ in $\mathbb{R}$, where ν_1 and ν_2 are allowed to be $-\infty$ and ∞ respectively. Using (2.1) as a working model for model-assisted estimation, the generalized entropy method can be formulated as minimizing

$$\sum_{i \in A} c_i \, G(\omega_i) \tag{4.1}$$

subject to (2.2) and

$$\sum_{i \in A} \omega_i \, g(d_i) \, c_i \; = \; \sum_{i=1}^{N} g(d_i) \, c_i, \tag{4.2}$$

where $g(\omega) = dG(\omega) / d\omega$ and c_i is to be determined. Constraint (4.2) plays the role of achieving the design consistency of the resulting calibration estimator and it is also called the debiasing constraint under model heterogeneity. Because the proposed calibration weighting uses a superpopulation model and achieves design consistency, it can be called model-assisted calibration using generalized entropy. Examples of generalized entropies and their debiasing calibration constraints can be found in Table 4.1.

Table 4.1
Examples of generalized entropies, $G(\omega)$, and the corresponding calibration covariates $g_i = g(\pi_i^{-1})$

Entropy	$G(\omega)$	$g_i = g(\pi_i^{-1})$	$f'(g_i)$	Domain $\mathcal{V}$
Squared Loss	ω^2	π_i^{-1}	1	$(-\infty, \infty)$
Empirical likelihood	$-\log \omega$	$-\pi_i$	π_i^{-2}	$(0, \infty)$
Exponential tilting	$\omega\,(\log(\omega) - 1)$	$-\log \pi_i$	π_i^{-1}	$(0, \infty)$
Hellinger distance	$-\sqrt{\omega}$	$-\pi_i^{1/2}$	$\pi_i^{-3/2}$	$(0, \infty)$
Rényi entropy $(\alpha \neq 0, -1)$	$\pm\omega^{\alpha+1}$	$\pm\pi_i^{-\alpha}$	$\pi_i^{\alpha-1}$	$(0, \infty)$

Notes: $\pm = \text{sgn}(\alpha + 1)\,\alpha$ is either + or − and determined so that the corresponding $G(\omega_i)$ is convex. Multiplicative constants are ignored. f' is the first-order partial derivative of $f = g^{-1}$.

To discuss asymptotic properties of the generalized calibration estimator, we assume a sequence of finite populations and samples as discussed in Isaki and Fuller (1982). The following theorem presents the $\sqrt{n}$-consistency of the generalized entropy calibration estimator.

Theorem 1. *Let $\hat{\omega}_i$ be obtained by minimizing (4.1) subject to (2.2) and (4.2). Under some regularity conditions stated in Appendix A, the resulting calibration estimator $\hat{Y}_{\text{gcal}} = \sum_{i \in A} \hat{\omega}_i \, y_i$ satisfies*

$$\hat{Y}_{\text{gcal}} \; = \; \hat{Y}_{\text{gcal}, \ell} + o_p\,(n^{-1/2} N),$$

where

$$\hat{Y}_{\text{gcal}, \ell} \; = \; \sum_{i=1}^{N} \mathbf{z}_i' \, \boldsymbol{\gamma}_q^* + \sum_{i \in A} \frac{1}{\pi_i} \left(y_i - \mathbf{z}_i' \, \boldsymbol{\gamma}_q^* \right), \tag{4.3}$$

$\mathbf{z}_i' = (\mathbf{x}_i', g\,(d_i)\,c_i)$ and $\boldsymbol{\gamma}_q^*$ is the probability limit of $\hat{\boldsymbol{\gamma}}_q$ given by

$$\hat{\boldsymbol{\gamma}}_q \; = \; \left(\sum_{i \in A} q_i^{-1} \, \mathbf{z}_i \, \mathbf{z}_i' \right)^{-1} \sum_{i \in A} q_i^{-1} \, \mathbf{z}_i \, y_i. \tag{4.4}$$

with $q_i = g'(d_i) c_i$.

By Theorem 1, the proposed calibration estimator is asymptotically equivalent to the debiased projection estimator using the augmented covariate $\mathbf{z}$:

$$\hat{Y}_{\text{d, proj}} = \sum_{i \in U} \mathbf{z}_i' \hat{\boldsymbol{\gamma}}_q + \sum_{i \in A} \frac{1}{\pi_i} \left(y_i - \mathbf{z}_i' \hat{\boldsymbol{\gamma}}_q \right).$$

Using the same argument for (3.5), we can express

$$\hat{Y}_{\text{d, proj}} = \sum_{i \in U} \mathbf{x}_i' \hat{\boldsymbol{\gamma}}_1 + \left(\frac{\sum_{i \in U} z_{2i}}{\sum_{i \in A} \pi_i^{-1} z_{2i}} \right) \sum_{i \in A} \frac{1}{\pi_i} \left(y_i - \mathbf{x}_i' \hat{\boldsymbol{\gamma}}_1 \right) \tag{4.5}$$

where $z_{2i} = g(d_i) c_i$,

$$\hat{\boldsymbol{\gamma}}_1 = \left(\sum_{i \in A} q_i^{-1} \tilde{\mathbf{z}}_{1i} \tilde{\mathbf{z}}_{1i}' \right)^{-1} \sum_{i \in A} q_i^{-1} \tilde{\mathbf{z}}_{1i} y_i$$

and $\tilde{\mathbf{z}}_{1i} = \mathbf{x}_i - z_{2i} \left(\sum_{i \in A} q_i^{-1} z_{2i}^2 \right)^{-1} \left(\sum_{i \in A} q_i^{-1} z_{2i} \mathbf{x}_i \right)$. When the sampling mechanism is informative, then the ratio adjustment term improves the efficiency of the bias-correction term in (4.5).

Note that Theorem 1 does not use the superpopulation model in (2.1) as an assumption. If the superpopulation model is indeed correct, then we obtain $\boldsymbol{\gamma}^* = (\boldsymbol{\beta}', 0)'$ and

$$\hat{Y}_{\text{greg}, \ell} = \sum_{i=1}^{N} \mathbf{x}_i' \boldsymbol{\beta} + \sum_{i \in A} \frac{1}{\pi_i} \left(y_i - \mathbf{x}_i' \boldsymbol{\beta} \right) + o_p \left(n^{-1/2} N \right).$$

In this case, the asymptotic variance of $\hat{Y}_{\text{gcal}}$ is equal to

$$V\left(\hat{Y}_{\text{greg}, \ell} \right) = V\left(\sum_{i=1}^{N} y_i \right) + E\left\{ \sum_{i=1}^{N} \left(\pi_i^{-1} - 1 \right) v_i \right\}, \tag{4.6}$$

which achieves the Godambe-Joshi lower bound of the anticipate variance under the model in (2.1). The optimal design minimizing the asymptotic variance (4.6) under a fixed sample size is achieved when

$$\pi_i \propto v_i^{1/2},$$

which was first established by Isaki and Fuller (1982).

If the superpopulation model is incorrect, Theorem 1 is still applicable and the asymptotic variance of $\hat{Y}_{\text{gcal}}$ is equal to

$$V\left(\hat{Y}_{\text{greg}, \ell} \right) = V\left(\sum_{i=1}^{N} y_i \right) + E\left\{ \sum_{i=1}^{N} \left(\pi_i^{-1} - 1 \right) \left(y_i - \mathbf{z}_i' \boldsymbol{\gamma}^* \right)^2 \right\}. \tag{4.7}$$

On the other hand, the classical calibration estimator $\hat{\theta}_{\text{DS}}$ of Deville and Särndal (1992) can be described as minimizing

$$Q(\boldsymbol{\omega}) = \sum_{i \in A} d_i \, G(\omega_i / d_i) \, c_i \tag{4.8}$$

subject to the calibration constraints in (2.1), where $d_i = \pi_i^{-1}$. Let $\hat{\omega}_i$ be the solution to the above optimization problem and $\hat{Y}_{\mathrm{DS}} = \sum_{i \in A} \hat{\omega}_i y_i$ be the resulting calibration estimator. Under some conditions on $G(\cdot)$, $\hat{Y}_{\mathrm{DS}}$ is asymptotically equivalent to the classical GREG estimator given by

$$\hat{Y}_{\mathrm{GREG}} = \sum_{i=1}^{N} \mathbf{x}_i' \hat{\boldsymbol{\beta}}_{\mathrm{GLS}} + \sum_{i \in A} \frac{1}{\pi_i} \left(y_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}}_{\mathrm{GLS}} \right)$$

where

$$\hat{\boldsymbol{\beta}}_{\mathrm{GLS}} = \left(\sum_{i \in A} d_i \mathbf{x}_i \mathbf{x}_i' / c_i \right)^{-1} \sum_{i \in A} d_i \mathbf{x}_i y_i / c_i.$$

Here, the subscript GLS denotes the generalized least squares method. Thus, ignoring the smaller order terms, satisfies

$$V\left(\hat{Y}_{\mathrm{DS}}\right) = V\left(\sum_{i=1}^{N} y_i \right) + E\left\{ \sum_{i=1}^{N} \left(\pi_i^{-1} - 1 \right) \left(y_i - \mathbf{x}_i' \boldsymbol{\beta}^* \right)^2 \right\}, \tag{4.9}$$

where $\boldsymbol{\beta}^*$ is the probability limit of $\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}$. Comparing (4.7) with (4.9), the additional covariate $g_i c_i$ in $\mathbf{z}_i$ can improve the prediction power for y_i. Thus, the proposed calibration estimator is more efficient than the classical calibration estimator when the superpopulation model is incorrect.

Remark 1. *The above theory is established for any c_i in the calibration problem using (4.1) and the choice of c_i is somewhat flexible. In practice, we can use either $c_i = v_i$ or use $c_i = v_i / g'(d_i)$. For the choice of $c_i = v_i / g'(d_i)$, we can obtain $q_i = v_i$ which leads to model-optimal regression estimation when the model variance v_i is correctly specified. Further investigation on the choice of c_i will be pursued in the future.*

For variance estimation, we can use the main result in Theorem 1 to obtain the following linearization variance estimator.

$$\hat{V}\left(\hat{Y}_{\mathrm{gcal}}\right) = \sum_{i \in A} \sum_{j \in A} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \frac{y_i - \mathbf{z}_i' \hat{\boldsymbol{\gamma}}_q}{\pi_i} \frac{y_j - \mathbf{z}_i' \hat{\boldsymbol{\gamma}}_q}{\pi_j} \tag{4.10}$$

where π_{ij} is the joint inclusion probability and $\hat{\boldsymbol{\gamma}}_q$ is defined in (4.4).

5. Model-based justification

The proposed method in Section 4 does not require that the superpopulation model in (2.1) is correctly specified. In this section, we will investigate statistical properties of the proposed estimator under the superpopulation model.

Under model (2.1), we can express

$$\begin{aligned} \hat{Y}_{\mathrm{gcal}} - Y &= \sum_{i \in A} \hat{\omega}_i y_i - \sum_{i=1}^{N} y_i \\ &= \sum_{i \in A} \hat{\omega}_i e_i - \sum_{i=1}^{N} e_i \end{aligned}$$

where the second equality holds by (2.2). Because e_i has zero expectation, we can expect that $E_\zeta(\hat{Y}_{\text{gcal}} - \hat{Y}) = 0$ under the model. Now, assuming $v_i = \mathbf{x}_i' \boldsymbol{\lambda}$ for some $\boldsymbol{\lambda}$, the model variance is

$$\begin{aligned} E\left\{\left(\hat{Y}_{\text{gcal}} - Y\right)^2\right\} &= E\left\{\sum_{i \in A} \hat{\omega}_i^2 v_i - \sum_{i=1}^{N} v_i\right\} \\ &= E\left\{\sum_{i \in A}\left(\hat{\omega}_i^2 - \hat{\omega}_i\right) v_i\right\} \end{aligned} \tag{5.1}$$

where the second equality holds by (2.2) and the fact that v_i is a part of $\mathbf{x}_i$. Note that the above variance is exact in the sense that the result does not require that the sample size be large.

Thus, to improve the finite-sample performance of the design-based variance estimator using the superpopulation model, we may employ the weighted residual technique of Särndal, Swenson and Wretman (1989) to obtain

$$\hat{V} = \sum_{i \in A} \sum_{j \in A} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \hat{\omega}_i \hat{e}_i \hat{\omega}_j \hat{e}_j \tag{5.2}$$

as a modified variance estimator of $\hat{Y}_{\text{gcal}}$. Under the design-based approach, $\hat{\omega}_i$ converges in probability to π_i^{-1} and the above variance estimator is asymptotically equivalent to the design-based variance estimator in (4.10). Also, if the superpopulation model is correct, then $\hat{e}_i$ will converge to e_i and we obtain

$$\begin{aligned} E\left\{\sum_{i \in A} \sum_{j \in A} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \hat{\omega}_i e_i \hat{\omega}_i e_j\right\} &= E\left\{\sum_{i \in A} \hat{\omega}_i^2 \left(1 - \pi_i\right) v_i\right\} \\ &\cong E\left\{\sum_{i \in A}\left(\hat{\omega}_i^2 - \hat{\omega}_i\right) v_i\right\} \end{aligned}$$

which provides a model-based justification. Therefore, we expect that the proposed variance estimator in (5.2) has a better performance in finite samples, especially when the regression model (2.1) is correct.

6. Computational details

A dual problem to the constrained problem in (4.1) is minimizing

$$Q_b(\boldsymbol{\gamma}) = \sum_{i \in A} c_i F(\mathbf{z}_i' \boldsymbol{\gamma} / c_i) - \sum_{i \in U} \mathbf{z}_i' \boldsymbol{\gamma}, \tag{6.1}$$

where F is the convex conjugate of G satisfying $F(\omega) = -G(g^{-1}(\omega)) + g^{-1}(\omega)\,\omega$. Let $f(\omega) = dF(\omega)/d\omega$. The solution to the dual problem can be expressed as

$$\omega_i^*(\boldsymbol{\gamma}) = f(\mathbf{z}_i' \boldsymbol{\gamma} / c_i), \quad i \in A.$$

where $f = g^{-1}$. Although (6.1) is a low-dimensional unconstrained problem, it requires a closed-form expression for F, which is not always possible. In this case, we can consider the primal problem of minimizing

$$Q_a(\boldsymbol{\omega}) = \sum_{i \in A} c_i\, G(\omega_i) + \sum_{k=1}^{p} \lambda_k \left(\sum_{i \in A} \omega_i\, x_{ik} - \sum_{i=1}^{N} x_{ik} \right)^2 + \lambda_{p+1} \left(\sum_{i \in A} \omega_i\, g_i\, v_i - \sum_{i=1}^{N} g_i\, v_i \right)^2 \qquad (6.2)$$

with $\lambda_k \to \infty$ for $k = 1, \ldots, p+1$. Since $Q_a(\boldsymbol{\omega})$ is a convex function, the unconstrained optimization in (6.2) is easier to handle than the constrained optimization. In R, we can use **optim** function with "L-BFGS-B" option to solve (6.2) with range constraints in the final weights. Furthermore, if $\mathbf{x}_i$ is high dimensional, we can relax the hard calibration by allowing some of λ_k bounded. This is directly related to soft calibration (Chambers, 1996; Guggemos and Tillé, 2010)

A dual problem to the unconstrained problem in (6.2) is minimizing

$$Q_b(\boldsymbol{\gamma}) = \sum_{i \in A} c_i F(\mathbf{z}_i' \boldsymbol{\gamma} / c_i) - \sum_{i \in U} \mathbf{z}_i' \boldsymbol{\gamma} + \boldsymbol{\gamma}' \operatorname{diag}(\lambda_1^{-1}, \ldots, \lambda_{p+1}^{-1})\, \boldsymbol{\gamma}, \qquad (6.3)$$

where F is the convex conjugate of $G.$ As $\boldsymbol{\lambda} \to \infty,$ the last term of (6.3) vanishes and (6.3) coincides with the dual problem of the exact calibration method in (4.1). A justification of the dual problem in (6.3) is presented in Appendix.

The proposed methodology can be implemented using the open-source R package `GECal` at https://CRAN.R-project.org/package=GECal. By suitably specifying a model that is associated with the calibration constraints, `GECal` performs various kinds of calibration weighting, including Deville and Särndal (DS) method and the proposed generalized entropy calibration (GEC) method. The package produces the calibration weights, survey estimates, and their standard errors for statistical inference.

7. Simulation study

We consider two simulation studies. One is based on a synthetic finite population and the other is based on a real survey data. In the first simulation, we know the true superpopulation model. Thus, we can evaluate the performance of the proposed methods under the correct model as well as under the mis-specified model. When a real data is used in the simulation, we do not know the true superpopulation model that generates the finite population. Thus, the second simulation study reflects a more realistic situation.

7.1 Simulation study one

In the first simulation, we generate a synthetic dataset to test our theory. A finite population $\{(\mathbf{x}_i', y_i): i = 1, \ldots, N\}$ of size $N =$ 5,000 are generated as follows: The auxiliary variables $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3})$'s were generated from $x_{ij} \sim N(2, 1)$ for $j = 1, 2, 3.$ The experiemental design for the simulation study can be described as a 2×2 factorial design with the following two factors.

1. Data generating model: linear versus nonlinear model.
2. Sampling mechanism: non-informative sampling versus informative sampling.

We consider two different data generating models for the study variable y_i. One is a linear regression model $y_i = 2 + x_{i1} + 2x_{i2} + 3x_{i3} + e_i$ and the other is a nonlinear model $y_i = 2.5 + x_{i1}^2 + \exp(x_{i2})/2 + e_i$ where $e_i \sim N(0, v_i)$ independently with $v_i = \exp(-x_{i1} + x_{i3})$. From each finite population, samples are repeatedly selected using a Poisson sampling whose inclusion probability is either $\pi_i = \text{expit}(-1 + 0.1\,x_{i1} - 0.1\,x_{i2} - 0.2\,x_{i3})$ for non-informative sampling or $\pi_i = \text{expit} - (1.25 + 0.1\,(-x_{i1} + e_i + x_{i1}e_i))$ for informative sampling. The inclusion probabilities are truncated above at 0.07 to avoid extremely large design weights. Both sampling design have expected sample size $E(n) \approx$ 1,000. The size of the Monte Carlo sample is $B =$ 1,000.

We consider the following point estimators.

- Hajék estimator
- Deville and Särndal (DS) estimator $\hat{Y}_{\text{DS}} = \sum_{i \in A} \omega_i y_i$, where $\hat{\omega}_i$ solves the optimization problem (4.8) subject to (2.1) using $c_i = 1$.
- The proposed generalized entropy calibration (GEC) estimator using $c_i = 1$.

For the DS and the GEC estimator, we used $G(\omega) = -\log \omega$ (EL), $G(\omega) = \omega \log \omega$ (ET), and $G(\omega) = \omega^2$ (SL). Note that the DS estimator using SL is exactly equal to the classical GREG estimator and the DS estimator using EL is equal to the pseudo empirical likelihood estimator of Wu and Rao (2006). Also, under Poisson sampling, the GEC estimator using EL is equal to the empirical likelihood estimator of Oguz-Alper and Berger (2016).

For variance estimation, we used the linearization variance estimator in (4.10) (Var1) and weighted residual variance estimator in (5.2) of Särndal et al. (1989) (Var2). We computed a relative bias of the variance estimator and the coverage rate of its 95% level confidence interval.

Table 7.1 presents the performance of the point estimators and their variance estimators under four scenarios. The t-statistics is computed by

$$T = \frac{E_{\text{MC}}(\hat{\theta}) - \theta}{\sqrt{V_{\text{MC}}(\hat{\theta})}} = \sqrt{B} \cdot \frac{B^{-1} \sum_{k=1}^{B} \hat{\theta}^{(k)} - \theta}{\left\{(B-1)^{-1} \sum_{k=1}^{B} (\hat{\theta}^{(k)} - \bar{\theta}_B)^2\right\}^{1/2}},$$

where $B =$ 1,000 is the size of the Monte Carlo sample and $\hat{\theta}^{(k)}$ denotes the realized value of $\hat{\theta}$ obtained from the k-th Monte Carlo sample. The T-statistic is a test statistic for testing $H_0: E(\hat{\theta}) = \theta$.

Overall, all the estimators considered revealed negligible bias. When the superpopulation model is linear and the sampling scheme is non-informative, the DS estimators and GEC estimators exhibit similar performance. However, the GEC estimators show smaller standard error than the DS estimators especially when the super-population model is non-linear or when the sampling scheme is informative. Under a linear model, the GEC estimators showed a dramatic decrease in standard error compared to the DS estimators under the informative sampling design. This is because there was a great increase in the prediction power of y_i's after adding the design covariate $g(d_i)$. When it comes to the variance estimation, the two variance estimators show similar performance.

Table 7.1
t-statistics and standard error (SE) of the point estimators and relative bias (RB) and 95% coverage rate (CR) of the variance estimators, based on $B=$ 1,000 Monte Carlo samples

		Point est		Var est1		Var est2	
		t	SE	RB	CR	RB	CR
Linear model with non-informative sampling							
	Hajék	0.09	593	0.089	0.96	0.089	0.96
EL	DS	0.19	283	0.011	0.95	0.011	0.95
	GEC	0.26	284	-0.014	0.94	-0.002	0.95
ET	DS	0.17	282	0.013	0.95	0.011	0.95
	GEC	0.19	282	0.001	0.94	0.004	0.95
SL	DS	0.17	282	0.017	0.95	0.013	0.95
	GEC	0.22	282	0.001	0.94	0.002	0.95
Linear model with informative sampling							
	Hajék	1.67	607	-0.087	0.93	-0.087	0.93
EL	DS	0.25	275	1.177	0.99	1.154	0.99
	GEC	1.33	121	-0.059	0.93	-0.019	0.93
ET	DS	0.22	274	1.184	0.99	1.158	0.99
	GEC	0.74	120	-0.022	0.93	-0.023	0.94
SL	DS	-0.55	337	0.448	0.97	0.438	0.97
	GEC	-0.84	173	-0.054	0.92	-0.059	0.93
Nonlinear model with non-informative sampling							
	Hajék	0.54	1284	0.036	0.95	0.036	0.95
EL	DS	-0.07	758	0.092	0.96	0.097	0.96
	GEC	0.15	750	0.008	0.95	0.027	0.95
ET	DS	-0.51	757	0.095	0.96	0.095	0.95
	GEC	-0.20	752	0.012	0.95	0.020	0.95
SL	DS	-0.78	766	0.070	0.95	0.066	0.95
	GEC	-0.31	760	0.004	0.94	0.010	0.95
Nonlinear model with informative sampling							
	Hajék	0.72	1391	-0.088	0.94	-0.088	0.94
EL	DS	1.02	746	0.192	0.97	0.199	0.97
	GEC	0.70	695	-0.017	0.95	0.002	0.95
ET	DS	0.48	746	0.191	0.97	0.195	0.97
	GEC	0.13	692	-0.006	0.95	0.006	0.95
SL	DS	-0.10	777	0.097	0.97	0.099	0.96
	GEC	-0.64	700	0.008	0.95	0.017	0.94

Note : EL = empirical likelihood; ET = exponential tilting; SL = squared loss; HD = Hellinger distance; DS = Deville and Särndal; GEC = generalized entropy calibration.

7.2 Simulation study two

In the second simulation study, we apply the proposed methods to the 1998 Survey of Mental Health Organizations (SMHO), which provides information on mental health care facilities and general hospital mental health services. This dataset, available in the `R` package `PracTools`, is originally survey data without population totals on auxiliary variables. For illustration, we treated the sampled data as a pseudo-population and repeatedly drew samples. We excluded nine observations with unusually high total expenditure of a health organization ($>10^8$\$), resulting in a final sample of 866 observations.

The total expenditure of a health organization (`EXPTOTAL`) is the study variable, which ranges from 141.36×10^3\$ to 93.11×10^6\$. An auxiliary variable related to the study variable is the end-of-year count of patients on the role (`EOYCNT`). The model R^2 is 0.1229, implying a weak positive linear relationship between `EOYCNT` and `EXPTOTAL`. The sampling scheme is a stratified Poisson sample with five strata, with a measure of size proportional to the squared root of the number of beds plus one $\left(\sqrt{\texttt{BEDS}+1}\right)$. The expected sample size is $E(n) = 50$. Similarly to the simulation study, we computed the Hajék estimator, Deville and Särndal (DS) estimator, and the proposed generalized entropy calibration (GEC) estimator. For DS and GEC estimators, three different entropies are considered: empirical likelihood (EL), exponential tilting (ET), and Hellinger distance (HD, $G(\omega) = -\sqrt{\omega}$). We did not incorporate the variance model $v_i = V(Y_i \mid \mathbf{x}_i)$ in weighting the generalized entropy. The Monte Carlo bias (Bias), standard error (SE), and root mean-squared-error (RMSE) of the point estimators are presented in Table 7.2.

As expected, all the estimators considered exhibited negligible design bias. A mild association between the study variable `EXPTOTAL` and the auxiliary variable `EOYCNT` led to the lower RMSE of the DS and GEC estimator compared to the Hajék estimator. However, the strong prediction power of the debiasing covariate $\left(g(\sqrt{\texttt{BEDS}+1})\right)$ on the study variable (`EXPTOTAL`) significantly reduces the variance of the GEC estimator compared to the DS estimator. Among the GEC estimators with three different entropies, HD entropy provided the smallest RMSE.

Table 7.2
Bias, SE, and RMSE of the point estimators using smho dataset $(\times 10^8)$

		Point est		
		Bias	**SE**	**RMSE**
	Hajék	0.79	15.86	15.88
EL	DS	1.45	14.81	14.88
	GEC	-0.56	13.22	13.23
ET	DS	1.38	14.74	14.80
	GEC	-1.41	13.26	13.33
HD	DS	1.41	14.78	14.85
	GEC	-0.86	13.02	13.05

Note : SE = standard error; RMSE = root mean-squared-error; EL = empirical likelihood; ET = exponential tilting; HD = Hellinger distance; DS = Deville and Särndal; GEC = generalized entropy calibration.

8. Concluding remarks

The generalized entropy calibration has been introduced as a new method for model-assisted calibration weighting. The proposed method is design consistent and is efficient when the underlying model is correct. The efficiency gain is significant over the method of Deville and Särndal (1992) when the sampling design is informative. However, the proposed method is applicable when the first-order inclusion probabilities are available throughout the finite population. Otherwise, the efficiency gain may not be achieved. An R-package, `GECal`, has been developed to implement the proposed method.

There are several possible extensions of the proposed method. The proposed method can be directly applicable to nonresponse weighting adjustment. Doubly robust (Kim and Haziza, 2014) or multiply robust (Chen and Haziza, 2017) calibration weighting method can be developed in this case. If the response mechanism is not ignorable, then a nonignorable nonresponse model can be considered. In this case, the empirical likelihood approach of (Qin, Leung and Shao, 2002; Liu and Fan, 2023) can be modified to solve the problem. Furthermore, the proposed method can be used to develop a method for data integration (Chen, Li and Wu, 2020). Such extensions will be investigated in the future.

Acknowledgements

We thank the Editor, the AE, and an anonymous referee for their constructive comments. The research of the first author was done during his visit to Seoul National University which was supported by Brain Pool program funded by the Ministry of Science and ICT through the National Research Foundation of Korea (RS-2023-00218474). His research was also partially supported by a grant from the U.S. National Science Foundation (2242820) and a grant from the U.S. Department of Agriculture's National Resources Inventory, Cooperative Agreement NR203A750023C006, Great Rivers CESU 68-3A75-18-504. Junyong Park was supported by Brain Pool program funded by the Ministry of Science and ICT through the National Research Foundation of Korea (RS-2023-00218474) and partially supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. 2020R1A2C1A01100526).

Appendix

A. Regularity conditions

We prove Theorem 1 under the following regularity conditions:

[A1] $G(\omega)$ is strictly convex and twice continuously differentiable in an open interval $\mathcal{V}$.

[A2] There exist positive constants $v_1, v_2 \in \mathcal{V}$ such that $v_1 < \pi_i < v_2$ for $i = 1, \ldots, N$.

[A3] Let π_{ij} be the joint inclusion probability of units i and j and $\Delta_{ij} = \pi_{ij} - \pi_i \pi_j$. Assume

$$\limsup_{N \to \infty} N \max_{i, j \in U:\, i \neq j} \left| \Delta_{ij} \right| < \infty.$$

[A4] Let $\mathbf{z}_i' = (\mathbf{x}_i', g(d_i)\, c_i)$. Assume $\mathbf{\Sigma}_{\mathbf{z}} = \lim_{N \to \infty} \sum_{i \in U} \mathbf{z}_i \mathbf{z}_i' / N$ exists and positive definite, the average 4th moment of $(y_i, \mathbf{x}_i')$ is finite such that $\limsup_{N \to \infty} \sum_{i=1}^{N} \left\| (y_i, \mathbf{x}_i') \right\|^4 / N < \infty$.

B. Proof of Theorem 1

Using Lagrange multiplier method, the optimization problem can be expressed as maximizing

$$Q_2(\boldsymbol{\omega}, \boldsymbol{\lambda}) = -\sum_{i \in A} G(\omega_i) c_i + \boldsymbol{\lambda}' \left(\sum_{i \in A} \omega_i \mathbf{z}_i - \sum_{i=1}^{N} \mathbf{z}_i \right).$$

Since

$$\frac{\partial}{\partial \omega_i} Q_2 = -g(\omega_i)\, c_i + \boldsymbol{\lambda}' \mathbf{z}_i,$$

the maximizer can be expressed as

$$\omega_i^*(\boldsymbol{\lambda}) = g^{-1}\left(\boldsymbol{\lambda}_1'\mathbf{x}_i / c_i + \lambda_2\, g_i\right)$$

where $g_i = g(d_i)$. Now, we can express

$$\hat{Y}_{\text{gcal}} = \hat{Y}_{\text{gcal}}(\hat{\boldsymbol{\lambda}}) = \sum_{i\in A}\omega_i^*(\hat{\boldsymbol{\lambda}})\, y_i$$

where $\hat{\boldsymbol{\lambda}}$ satisfies (2.1), we can express

$$\begin{aligned}\hat{Y}_{\text{gcal}} &= \sum_{i\in A}\omega_i^*(\hat{\boldsymbol{\lambda}})\, y_i + \underbrace{\left(\sum_{i=1}^N \mathbf{z}_i - \sum_{i\in A}\omega_i^*(\hat{\boldsymbol{\lambda}})\,\mathbf{z}_i\right)'}_{=0}\boldsymbol{\gamma}\\ &:= \hat{Y}_\ell(\hat{\boldsymbol{\lambda}}, \boldsymbol{\gamma}).\end{aligned}$$

The asymptotic existence and uniqueness of $\hat{\boldsymbol{\lambda}}$ that satisfies (2.1) can be justified following Kwon et al. (2024) under Conditions [A1] – [A4]. Let $\boldsymbol{\lambda}^*$ be the probability limit of $\hat{\boldsymbol{\lambda}}$. Since $\hat{\boldsymbol{\lambda}}$ satisfies (2.1), $\boldsymbol{\lambda}^*$ should satisfy

$$\underbrace{E\left\{\sum_{i\in A}\omega_i^*(\boldsymbol{\lambda}^*)\,\mathbf{z}_i \mid \mathcal{F}_N\right\}}_{=\sum_{i=1}^N \pi_i\,\omega_i^*(\boldsymbol{\lambda}^*)\,\mathbf{z}_i} = \sum_{i=1}^N \mathbf{z}_i,$$

which implies that

$$\omega_i^*(\boldsymbol{\lambda}^*) = \pi_i^{-1} = d_i$$

or

$$g^{-1}\left(\mathbf{x}_i'\boldsymbol{\lambda}_1^* / c_i + g(d_i)\,\lambda_2^*\right) = d_i.$$

Since $g(\cdot)$ is one-to-one, we get $\boldsymbol{\lambda}_1^* = \mathbf{0}$ and $\lambda_2^* = 1$.

Now, to obtain the linearization, we use the technique of Randles (1982) as discussed in Kim and Rao (2012). That is, we wish to find $\boldsymbol{\gamma}^*$ such that

$$E\left\{\frac{\partial}{\partial\boldsymbol{\lambda}}\hat{Y}_\ell(\boldsymbol{\lambda}^*, \boldsymbol{\gamma})\right\} = 0$$

holds at $\boldsymbol{\gamma} = \boldsymbol{\gamma}^*$. Now, since

$$\frac{\partial}{\partial\boldsymbol{\lambda}}\hat{Y}_\ell(\boldsymbol{\lambda}^*, \boldsymbol{\gamma}) = \sum_{i\in A}\frac{1}{g'\{g^{-1}(\mathbf{z}_i'\boldsymbol{\lambda}^* / c_i)\}}\left(y_i - \mathbf{z}_i'\boldsymbol{\gamma}\right)\mathbf{z}_i / c_i,$$

we obtain

$$\boldsymbol{\gamma}^* = \left(\sum_{i=1}^N \frac{\pi_i}{g'(d_i)\, c_i}\mathbf{z}_i\mathbf{z}_i'\right)^{-1}\sum_{i=1}^N \frac{\pi_i}{g'(d_i)\, c_i}\mathbf{z}_i\, y_i.$$

Therefore, we obtain (4.3).

C. Proof of (6.3)

Note that (6.2) is equivalent to solving

$$\underset{\boldsymbol{\omega}\in\mathbb{R}^n,\, \mathbf{u}\in\mathbb{R}^{p+1}}{\arg\min} \sum_{i\in A} c_i\, G(\omega_i) \; + \; \mathbf{u}'\,\mathrm{diag}\,(\lambda_1, \ldots, \lambda_{p+1})\,\mathbf{u} \tag{C.1}$$

subject to

$$\mathbf{u} \; = \; \sum_{i\in A} \omega_i\, \mathbf{z}_i - \sum_{i\in U} \mathbf{z}_i.$$

The Lagrangian of (C.1) is

$$\mathcal{L}(\boldsymbol{\omega}, \mathbf{u}, \boldsymbol{\gamma}) \; = \; \sum_{i\in A} c_i\, G(\omega_i) + \mathbf{u}'\,\mathbf{diag}(\boldsymbol{\lambda})\,\mathbf{u} + \boldsymbol{\gamma}'\left(\mathbf{u} - \left(\sum_{i\in A} \omega_i\, \mathbf{z}_i - \sum_{i\in U} \mathbf{z}_i\right)\right) \tag{C.2}$$

for $\boldsymbol{\omega}$, $\mathbf{u}$, and $\boldsymbol{\gamma} \in \mathbb{R}^{p+1}$. Differentiating with respect to $\boldsymbol{\omega}$ and $\mathbf{u}$ gives

$$\frac{\partial\mathcal{L}}{\partial\boldsymbol{\omega}} = \begin{pmatrix} c_1 g(\omega_1) \\ \vdots \\ c_n g(\omega_n) \end{pmatrix} - \begin{pmatrix} \mathbf{z}_1'\,\boldsymbol{\gamma} \\ \vdots \\ \mathbf{z}_n'\,\boldsymbol{\gamma} \end{pmatrix}$$

and

$$\frac{\partial\mathcal{L}}{\partial\mathbf{u}} \; = \; 2\,\mathrm{diag}\,(\boldsymbol{\lambda})\,\mathbf{u} + \boldsymbol{\gamma}.$$

Plugging the above equations to (C.2) gives

$$-Q_b(\boldsymbol{\gamma}) \; = \; \sum_{i\in A} \Big[G\big(f\,(\mathbf{z}_i'\,\boldsymbol{\gamma}) \big) - f\,(\mathbf{z}_i'\,\boldsymbol{\gamma})\,\mathbf{z}_i'\boldsymbol{\gamma} \Big] + \sum_{i\in U} \mathbf{z}_i'\boldsymbol{\gamma} - \boldsymbol{\gamma}'\mathrm{diag}\,(\boldsymbol{\lambda})^{-1}\boldsymbol{\gamma}.$$

References

Breidt, F.J., and Opsomer, J.D. (2017). Model-assisted survey estimation with modern prediction techniques. *Statistical Science*, 32, 190-205.

Chambers, R. (1996). Robust case-weighting for multipurpose establishment surveys. *Journal of Official Statistics*, 12, 3-32.

Chen, S., and Haziza, D. (2017). Multiply robust imputation procedures for the treatment of item nonresponse in surveys. *Biometrika*, 104, 439-453.

Chen, Y., Li, P. and Wu, C. (2020). Doubly robust inference with non-probability survey samples. *Journal of the American Statistical Association*, 115, 2011-2021.

Deville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418), 376-382.

Firth, D., and Bennett, K.E. (1998). Robust models in probability sampling. *Journal of the Royal Statistical Society: Series B*, 60, 3-21.

Fuller, W.A. (2002). Regression estimation for survey samples. *Survey Methodology*, 28(1), 5-23. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2002001/article/6408-eng.pdf.

Fuller, W.A. (2009). *Sampling Statistics*. New York: John Wiley & Sons, Inc.

Gneiting, T., and Raftery, A.E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477), 359-378.

Guggemos, F., and Tillé, Y. (2010). Penalized calibration in survey sampling: Design-based estimation assisted by mixed models. *Journal of Statistical Planning and Inference*, 140, 3199-3212.

Isaki, C.T., and Fuller, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.

Kim, J.K., and Haziza, D. (2014). Doubly robust inference with missing data in survey sampling. *Statistica Sinica*, 24, 375-394.

Kim, J.K., and Park, M. (2010). Calibration estimation in survey sampling. *International Statistical Review*, 78(1), 21-39.

Kim, J.K., and Rao, J.N.K. (2012). Combining data from two independent surveys: A model-assisted approach. *Biometrika*, 99, 85-100.

Kott, P.S. (2015). Calibration weighting in survey sampling. *WIREs Computational Statistics*, 8, 39-53.

Kwon, Y., Kim, J.K. and Qiu, Y. (2024). Debiased calibration estimation using generalized entropy in survey sampling. Unpublished manuscript (arXiv:2404.01076).

Liu, Y., and Fan, Y. (2023). Biased-sample empirical likelihood weighting for missing data problems: An alternative to inverse probability weighting. *Journal of the Royal Statistical Society: Series B*, 85, 67-83.

Oguz-Alper, M., and Berger, Y.G. (2016). Modelling complex survey data with population level information: An empirical likelihood approach. *Biometrika*, 103, 447-459.

Pfeffermann, D., and Sverchkov, M. (2009). Inference under informative sampling. *Handbook of Statistics,* Amsterdam: Elsevier, 29B, 455-487.

Qin, J., Leung, D. and Shao, J. (2002). Estimation with survey data under nonignorable nonresponse or informative sampling. *Journal of the American Statistical Association*, 97(457), 193-200.

Randles, R.H. (1982). On the asymptotic normality of statistics with estimated parameters. *The Annals of Statistics*, 10(2), 462-474.

Särndal, C.-E., Swenson, B. and Wretman, J. (1989). The weighted residual technique for estimating the variance of the general regression estimator. *Biometrika*, 76, 527-537.

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

Särndal, C.-E., and Wright, R.L. (1984). Cosmetic form of estimators in survey sampling. *Scandinavian Journal of Statistics*, 11, 146-156.

Wu, C., and Rao, J.N.K. (2006). Pseudo-empirical likelihood ratio confidence intervals for complex surveys. *The Canadian Journal of Statistics*, 34, 359-375.

Wu, C., and Sitter, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96, 185-193.