Debiased Calibration for Generalized Two-Phase Sampling

Caleb Leedy

April 15, 2024

1 Introduction

Combining information from several sources is an important practical problem. (CITEME) We want to incorporate information from external data sources to reduce the bias in our estimates or improve the estimator's efficiency. For many problems, the additional information consist of summary statistics with standard errors. The goal of this project is to incorporate external information with existing data to create more efficient estimators using calibration weighting.

To model this scenaro, we formulate the problem as a generalized two-phase sample where the first phase sample consists of data from multiple sources. The second phase sample contains our existing data. To motivate this setup, we consider the following approach: first, we consider the classical two-phase sampling setup where the second phase sample is a subset of the first phase sample; then, we extend this setup to consider non-nested two-phase samples; and finally, we consider the more general approach of having multiple sources as the first phase sample.

2 Topic 1: Classical Two-Phase Sampling

2.1 Background

Consider a finite population of size N containing elements (X_i, Y_i) where an initial (Phase 1) sample of size n is selected and X_i is observed. Then from the Phase 1 sample of elements,

a (Phase 2) sample of size $r < n$ is selected and Y_i is observed. This is two-phase sampling (See Fuller (2009), Kim (2024) for general references.) The goal of two-phase sampling is to construct an estimator of Y not only using the observed information from the Phase 2 sample but also incorporating the extra auxiliary information of X from the Phase 1 sample, and the challenge is doing this efficiently.

An easy-to-implement unbiased estimator in the spirit of a Horvitz-Thompson (HT) estimator (Horvitz and Thompson (1952), Narain (1951)) is the π^*-estimator. Let $\pi_i^{(2)}$ be the response probability of element i being observed in the Phase 2 sample. Then, allowing the elements in the Phase 1 sample to be represented by A_1 and the elements in the Phase 2 sample to be denoted as A_2, if we define $\pi_{2i|1} = \sum_{A_2: i \in A_2} \Pr(A_2 \mid A_1)$ and $\pi_{1i} = \sum_{A_1: i \in A_1} \Pr(A_1)$ then,

$$\pi_i(A_1) = \pi_{2i|1}\pi_{1i}.$$

I agree that having the superscript (1) and (2) is unnecessary. I have removed them. I have also modified the previous equation. I agree that the invariance condition does not hold for two-phase sampling, but if we consider π_i as a function of A_1 then I believe this holds. Is this correct? Should we not have $\pi_i = \pi_{1i}\pi_{2i|1}(A_1)$ because this is how conditional distributions work? (Also, I thought that this is equivalent to the reverse sampling idea of Fey (1992), Shao and Steel (1999), and Kim et al. (2006).) I would love to discuss this with you.

This means that we can define the π^*-estimator as the following design unbiased estimator:

$$\hat{Y}_{\pi^*} = \sum_{i \in A_2} \frac{y_i}{\pi_{2i|1}\pi_{1i}}.$$

While unbiased (see Kim (2024)), the π^*-estimator does not account for the additional information contained in the auxiliary Phase 1 variable X. The two-phase regression estimator $\hat{Y}_{reg,tp}$ does incorporate information for X by using the estimate $\hat{X}_1$ from the Phase 1 sample. This is how we can leverage the external information $\hat{X}_1$ to improve the initial

π^*-estimator in the second phase sample. The two-phase regression estimator has the form,

$$\hat{Y}_{reg,tp} = \sum_{i \in A_1} \frac{1}{\pi_{1i}} x_i \hat{\beta}_2 + \sum_{i \in A_2} \frac{1}{\pi_{1i}\pi_{2i|1}} (y_i - x_i \hat{\beta}_2)$$

where for $q_i = q(x_i)$ and is a function of x_i,

$$\hat{\beta}_2 = \left(\sum_{i \in A_2} \frac{x_i x_i'}{\pi_{1i} q_i} \right)^{-1} \sum_{i \in A_2} \frac{x_i y_i}{\pi_{1i} q_i}.$$ [1] [2]

The regression estimator is the minimum variance design consistent linear estimator which is easily shown to be the case because $\hat{Y}_{reg,tp} = \sum_{i \in A_2} \hat{w}_{2i} y_i / \pi_{1i}$ where

$$\hat{w}_{2i} = \text{argmin}_w \sum_{i \in A_2} (w_{2i} - \pi_{2i|1}^{-1})^2 q_i \text{ such that } \sum_{i \in A_2} w_{2i} x_i / \pi_{1i} = \sum_{i \in A_1} x_i / \pi_{1i}.$$

This means that $\hat{Y}_{reg,tp}$ is also a calibration estimator. The idea that regression estimation is a form of calibration was extended by Deville and Sarndal (1992) to consider loss functions other than just squared loss. They generalized the loss function to minimize $\sum_i G(w_i, d_i) q_i$ for weights w_i and design-weights d_i where $G(\cdot)$ is non-negative, strictly convex function with respect to w, defined on an interval containing d_i, with $g(w_i, d_i) = \partial G / \partial w$ continuous. [3] This generalization includes empirical likelihood estimation, and maximum entropy estimation among others. The variance estimation is based on a linearization that shows that minimizing the generalized loss function subject to the calibration constraints is asymptotically equivalent to a regression estimator.

While this generalization is useful to an analyst who may want different properties of their estimator from maximum entropy estimation rather than the minimal squared loss,

[1] Caleb, we do not have to use $\pi_{2i|1}^{(2)}$ in computing $\hat{\beta}_2$. Please check my book. Thanks for the check. I checked it and I must have gotten π_{1i} and π_{2i} confused.

[2] I was initially thinking of using $\hat{\beta}_2$ without q_i to be more consistent with how I defined the two-phase regression estimator; however, I see your point about using q_i to connect this estimator with the superpopulation model.

[3] The Deville and Sarndal (1992) paper considers regression estimators for a single phase setup, which we apply to our two-phase example.

unless $\pi_{2i|1}^{-1} = x_i'a$ for some a, the estimator is not design consistent. [4] [5]

Furthermore, at a conceptual level, the regression estimator has a nice feature that its two terms can be thought about as minimizing variance and bias correction,

$$\hat{Y}_{reg,tp} = \underbrace{\sum_{i\in A_1} \frac{x_i\hat{\beta}_2}{\pi_{1i}}}_{\text{Minimizing the variance}} + \underbrace{\sum_{i\in A_2} \frac{1}{\pi_{1i}\pi_{2i|1}}(y_i - x_i\hat{\beta}_2)}_{\text{Bias correction}}.$$

The Deville and Sarndal (1992) method incorporates the design weights into the loss function, which is the part minimizing the variance. Instead, there is a desire to get design consistency from the calibration term. In Kwon et al. (2024), the authors show that for a generalized entropy function $G(w)$, including a term of $g(\pi_{2i|1}^{-1})$ into the calibration for $g = \partial G/\partial w$ not only creates a design consistent estimator, but it also has better efficiency than the generalized regression estimators of Deville and Sarndal (1992).

The method of Kwon et al. (2024) requires known calibration levels (no uncertainty) for the finite population. It does not handle the two-phase setup where we need to estimate the finite population total of x from the Phase 1 sample. Hence, we extend this method to have a valid variance estimator when including estimated Phase 1 weights.

Methodology

We follow the approach of Kwon et al. (2024) for the debiased calibration method. We consider maximizing the generalized entropy Gneiting and Raftery (2007),

$$H(w) = -\sum_{i\in A_2} \frac{1}{\pi_{1i}} G(w_{2i})q_i \tag{1}$$

where $G : \mathcal{V} \to \mathbb{R}$ is strictly convex, differentiable function subject to the constraints:

$$\sum_{i\in A_2} \frac{x_i w_{2i}}{\pi_{1i}} = \sum_{i\in A_1} \frac{x_i}{\pi_{1i}} \tag{2}$$

[4] I do not understand this part. I changed the notation so that it would not be the unused column space notation. This is from Equation (11.17) of Kim (2024).

[5] I think that I am confused about the purpose of having the assumption from Equation (11.17) in your sampling book. Could we discuss this?

and

$$\sum_{i \in A_2} \frac{g(\pi_{2i|1}^{-1}) w_{2i} q_i}{\pi_{1i}} = \sum_{i \in A_1} \frac{g(\pi_{2i|1}^{-1}) q_i}{\pi_{1i}}. \tag{3}$$

The first constraint is the existing calibration constraint and the second ensures that design consistency is achieved. Here, $g(w) = \partial G / \partial w$. The original method of Kwon et al. (2024) only considered having finite population quantities on the right hand side of 2. Let $z_i = (x_i, g(\pi_{2i|1}^{-1})$. To solve this minimization problem, we use the convex conjugate function $F(w) = -G(g^{-1}) + g^{-1}(w)w$, and instead find

$$\hat{\lambda} = \text{argmin}_{\lambda \in \Lambda_A} \sum_{i \in A_2} F(w_i) + \lambda_1^T \sum_{i \in A_1} \frac{z_i}{\pi_{2i|1}}$$

for $\Lambda_A = \{\lambda : \lambda^T z_i \in g(\mathcal{V}) \text{ for all } i \in A_2\}$.

Theoretical Results

We prove two results in this section is the same spirit as Kwon et al. (2024). First, we state the assumptions of our analysis then we give the results.

(A1) $G(w)$ is strictly convex and twice differentiable in an interval $\mathcal{V} \in \mathbb{R}$,

(A2) There exists $c_1, c_2 \in \mathcal{V}$ with $c_1, c_2 > 0$ such that $c_1 < \pi_{2i|1} < c_2$ for $i = 1, \ldots, n_2$, and that $c_1 < \pi_{1i} < c_2$ for $i = 1, \ldots, n_1$,

(A3) If $\pi_{2ij|1}$ is joint inclusion probability of elements i and j in A_2, the Phase 2 sample and $\Delta_{2ij|1} = \pi_{2ij|1} - \pi_{2i|1}\pi_{2j|1}$ then,

$$\limsup_{n \to \infty} \max_{i,j \in U : i \neq j} |\Delta_{2ij|1}| < \infty,$$

and if π_{1ij} is the joint inclusion probability of elements i and j in A_1, the Phase 1 sample and $\Delta_{1ij} = \pi_{1ij} - \pi_{1i}\pi_{1j}$ then,

$$\limsup_{n \to \infty} \max_{i,j \in U : i \neq j} |\Delta_{1ij}| < \infty,$$

(A4) Assume that $\Sigma_z = \lim_{N\to\infty} \sum_{i\in U} z_i z_i^T/N$ exists and is positive definite, the average fourth moment of (y_i, x_i^T) is finite ($\limsup_{N\to\infty} \sum_{i\in U} ||(y_i, x_i^T)||^4/N < \infty$), and $\Gamma(\lambda) = \lim_{N\to\infty} \sum_{i\in U} f'(\lambda^T z_i) z_i z_i^T/N$ exists in a neighborhood around $\lambda_0 = (\mathbf{0}, 1)$.

Theorem 1 (Design Consistency)**.** *Suppose that Conditions (A1) - (A4) hold. Then the solution $\hat{w}$ to Equation 1 subject to Equation 2 and Equation 3 exists and is unique with probability approaching 1. Furthermore, the estimator $\hat{\theta}_{DCE} = \sum_{i\in A_2} \frac{\hat{w}_i y_i}{\pi_{1i}}$ satisfies*

$$\hat{\theta}_{DCE} = \hat{\theta}_{DC} + o_p(n_2^{-1})$$

where $\hat{\theta}_{DC}$ is the debiased calibration estimator of Kwon et al. (2024).

Proof. This proof follows the proof of Kwon et al. (2024) in a straightforward manner. There are two main modifications that need to be made. The first is that the proof needs to be written for two-phase sampling with n_1 and n_2 instead of a single n. The second is that we need to ensure that the result still holds with an estimated population constraint that is $O_p(n_1^{-1/2})$ consistent.

With the Conditions (A1) - (A4) written with respect the the two-phase sample, the first challenge is done. The second step largely holds from the existing proof of Kwon et al. (2024) with modifications such as

$$\hat{Q}(\lambda) = n_2^{-1} \sum_{i\in A_2} F(\lambda^T z_i^*) - n_1^{-1} \lambda^T \sum_{i\in A_1} z_i^*$$

ensuring that our extension holds.

Dr. Kim, I know that I need to write up this proof in more detail, but I am unsure about the level of granularity that is required now. I can basically rewrite everything from the Appendix of Kwon et al. (2024) with the two-phase modification to show this result, but I don't know if I should yet. □

We need the following additional assumptions to prove asymptotic normality.

I need to think about the assumptions that we need in two-phase sampling.

(B1)

(B2) The HT estimator is asymptotically normal under the sampling design in the sense that

Theorem 2 (Asymptotic Normality)**.** *I need to write out more but we should have something like this:*

$$V(\hat{\theta}_{DCE} = Var(\hat{\theta}_{DC}) + \hat{\gamma}_{[1:3]}^{T} V(\mathbf{x}))\hat{\gamma}_{[1:3]} + Cov(\hat{\theta}_{DC}, \mathbf{x}\hat{\gamma}_{[1:3]}]).$$

Dr. Kim, I am not convinced that $\text{Cov}(\hat{\theta}_{DC}, \mathbf{x}\hat{\gamma}_{[1:3]}) = 0$. The reason that this covariance holds in Fuller (2009) Section 3.3.3 is that the estimating equation requires $\sum_i x_i \hat{e}_i = 0$, which means that $\text{Cov}(\hat{e}, x_i) = 0$. However, we do not have this requirement. Instead, I see this more like a nested model where we have a group level error that aggregates the individual level errors. The problem with the current simulation that I cannot tell if this covariance should be zero because the variance term $\hat{\gamma}'_{[1:3]} V(\mathbf{x}) \hat{\gamma}_{[1:3]}$ dominates the variance estimation. I will change the simulation parameters to try to see if I can get a better test of this.

Simulation Studies

We run a simulation testing the proposed method. In this approach we have the following simulation setup:

$$
\begin{aligned}
X_{1i} &\overset{ind}{\sim} N(2,1) \\
X_{2i} &\overset{ind}{\sim} Unif(0,4) \\
X_{3i} &\overset{ind}{\sim} N(0,1) \\
X_{4i} &\overset{ind}{\sim} Unif(0.1,0.9) \\
\varepsilon_i &\overset{ind}{\sim} N(0,1) \\
Y_i &= 3X_{1i} + 2X_{2i} + \varepsilon_i \\
\pi_{1i} &= \Phi_3(-x_{3i} - 2) \\
\pi_{2i|1} &= x_{4i}.
\end{aligned}
$$

where Φ_3 is the CDF of a t-distribution with 3 degrees of freedom. This is a two-phase extension of the setup in Kwon et al. (2024). We consider a finite population of size $N = 10,000$ with both the Phase 1 and Phase 2 sampling occuring under Poisson (Bernoulli) sampling. This yields a Phase 1 sample size of $E[n_1] \approx 1100$ and a Phase 2 sample size of $E[n_2] \approx 550$. In the Phase 1 sample, we observe (X_1, X_2) while in the Phase 2 sample we observe (X_1, X_2, Y). This simulation does not deal with model misspecification, and we compare the proposed method for the parameter $\bar{Y}_N$ with four approaches:

1. π^*-estimator: $\hat{Y}_{\pi^*} = N^{-1} \sum_{i \in A_2} \frac{y_i}{\pi_{1i}\pi_{2i|1}}$,
2. Two Phase Regression estimator (TP-Reg): $\hat{Y}_{reg} = \sum_{i \in A_1} \frac{\mathbf{x}_i'\hat{\beta}}{\pi_{1i}} + \sum_{i \in A_2} \frac{1}{\pi_{1i}\pi_{2i|1}}(y_i - \mathbf{x}_i'\hat{\beta})$ where $\hat{\beta} = \left(\sum_{i \in A_2} \mathbf{x}_i\mathbf{x}_i'\right)^{-1} \sum_{i \in A_2} \mathbf{x}_i y_i$ and $\mathbf{x}_i = (x_{1i}, x_{2i})^T$, Dr. Kim, should I modify the simulation so that the regression estimator also includes $g(\pi_{2i|1}^{-1})$ as a covariate?
3. Debiased Calibration with Population Constraints (DC-Pop): This is the method from Kwon et al. (2024) with the true population level constraints, and
4. Debiased Calibration with Estimated Population Constraints (DC-Est): This is the proposed method with the Phase 1 sample being used to estimate the population level constraints.

In addition to estimating the mean parameter $\bar{Y}_N$, we also construct variance estimates $\hat{V}(\hat{Y})$ for each estimate $\hat{Y}$. For each approach we have the following variance estimate[6]:

Estimator	Variance	Notes
π^*	$N^{-2}\sum_{i\in A_2}\left(\pi_{2i\|1}^{-2}-\pi_{2i\|1}^{-1}\right)y_i^2$	
TP-Reg	$N^{-2}\left(\sum_{i\in A_1}\left(\pi_{1i}^{-2}-\pi_{1i}^{-1}\right)\eta_i^2+\sum_{i\in A_2}\frac{1}{\pi_{1i}\pi_{2i\|1}}(\pi_{2i\|1}^{-1}-1)(y_i-\mathbf{x}_i'\hat{\beta})^2\right)$	$\eta_i=\mathbf{x}_i\hat{\beta}+\frac{\delta_{2i}}{\pi_{2i\|1}}(y_i-\mathbf{x}_i\hat{\beta})$
DC-Pop	$(Y-\mathbf{z}^T\hat{\gamma})^T\Pi(Y-\mathbf{z}^T\hat{\gamma})$	$\Pi=\text{diag}(1-(\pi_1\pi_{2\|1})^{-1})\cdot\frac{\hat{w}^2}{N^2}$
DC-Est	$(Y-\mathbf{z}^T\hat{\gamma})^T\Pi(Y-\mathbf{z}^T\hat{\gamma})+\hat{\gamma}_{[1:3]}^T V(\mathbf{x})\hat{\gamma}_{[1:3]}/N^2$	

Figure 1: This table gives the formulas for each variance estimator used in this simulation.

We run this simulation 1000 times for each of these methods and compute the Bias ($E[\hat{Y}]-\bar{Y}_N$), the RMSE ($\sqrt{\text{Var}(\hat{Y}-\bar{Y}_N)}$), a 95% empirical confidence interval ($\sum_{b=1}^{1000}|\hat{Y}^{(b)}-\bar{Y}_N|\leq\Phi(0.975)\sqrt{\hat{V}(\hat{Y}^{(b)})^{(b)}}$), and a T-test that assesses the unbiasedness of each estimator. The results are in Figure 2.

Est	Bias	RMSE	EmpCI	Ttest
π^*	0.029	0.797	0.933	1.146
TP-Reg	0.003	0.457	0.957	0.227
DC-Pop	0.007	0.044	0.960	5.143
DC-Est	0.009	0.450	0.959	0.643

Figure 2: This table shows the results of the simulation study. It displays the Bias, RMSE, empirical 95% confidence interval, and a t-statistic assessing the unbiasedness of each estimator for the estimators: π^*, TP-Reg, DC-Pop, and DC-Est.

Topic 2: Non-nested Two-Phase Sampling

(Materials in Section 11.4 can be used here. I have copy-and-pasted the textbook materials below. Please modify them.) I will do this shortly.

In contrast to the classical two-phase sampling framework, non-nested two-phase sampling involves two independent surveys conducted on the same target population. The key

[6]These variance estimates use the fact that we have Poisson sampling for both phases in the simulation.

distinction is that the two samples, denoted as A_1 and A_2, are drawn independently rather than sequentially. Table 1 presents the data structure for non-nested two-phase sampling.

In the non-nested two-phase sampling, a large probability sample A_1 is drawn from a finite population, collecting only the $\mathbf{x}$ variable, and a smaller sample A_2 is drawn from the same population, providing information on both the y and $\mathbf{x}$ variables. It is assumed that the observed variable x is comparable in the two surveys. **?** formally addressed this non-nested two-phase sampling problem and **?** extended the idea further to develop regression estimation combining information from multiple surveys. **?** considered the non-nested two-phase sampling in the context of mass imputation combining two independent surveys at the population and domain levels.

Table 1: Data Structure for non-nested two-phase sampling

Sample	X	Y
A_1	✓	
A_2	✓	✓

To illustrate the non-nested two-phase sampling approach, let's consider the data structure shown in Table 1. This setup involves two independent samples, A_1 and A_2, drawn from the same target population.

From these two samples, we can compute two unbiased estimators of the population total $\mathbf{X} = \sum_{i=1}^{N} \mathbf{x}_i$ for the auxiliary variable x: $\hat{\mathbf{X}}_1 = \sum_{i \in A_1} \pi_{1i}^{-1} \mathbf{x}_i$ and $\hat{\mathbf{X}}_2 = \sum_{i \in A_2} \pi_{2i}^{-1} \mathbf{x}_i$. Here, π_{1i} and π_{2i} represent the inclusion probabilities for samples A_1 and A_2, respectively.

Both $\hat{\mathbf{X}}_1$ and $\hat{\mathbf{X}}_2$ are unbiased estimators of the population total $\mathbf{X}$ under the respective sampling designs. The availability of these two unbiased estimators is a key feature of the non-nested two-phase sampling design, as it provides opportunities for developing enhanced estimation procedures combining information from different sources.

We can construct a combined estimator of X, denoted as $\widehat{\mathbf{X}}_c$, as follows:

$$\widehat{\mathbf{X}}_c = W\hat{\mathbf{X}}_1 + (I - W)\hat{\mathbf{X}}_2, \tag{4}$$

where W is a $p \times p$ symmetric matrix of constants, and $p = \dim(\mathbf{x})$ is the dimension of the auxiliary variable x. The optimal choice of the matrix W can be determined using the

Generalized Least Squares (GLS) method. However, other choices of W can also be used. The key idea is to leverage the information from these two independent surveys to obtain a more accurate and efficient estimator of the population total X for the auxiliary variable x, compared to using only one of the surveys alone.

Using the combined estimator $\widehat{\mathbf{X}}_{\mathrm{c}}$ in (4), we can construct the following projection estimator:

$$\widehat{Y}_p = \widehat{\mathbf{X}}_{\mathrm{c}}' \hat{\boldsymbol{\beta}}_q \tag{5}$$

where the regression coefficient estimator $\hat{\boldsymbol{\beta}}_q$ is defined as

$$\hat{\boldsymbol{\beta}}_q = \left(\sum_{i \in A_2} \mathbf{x}_i \mathbf{x}_i' / q_i \right)^{-1} \sum_{i \in A_2} \mathbf{x}_i y_i / q_i.$$

The choice of q_i in the regression coefficient estimator is somewhat arbitrary. Two possible choices are:

1. Using the model variance under a regression superpopulation model.
2. Using $q_i = \pi_{2i}^{-2} - \pi_{2i}^{-1}$ to compute the design-optimal regression estimator under Poisson sampling.

The key idea is that by using the combined estimator $\widehat{\mathbf{X}}_c$ in the projection estimator $\widehat{Y}_p$, we can leverage the information from both the A1 and A2 samples to obtain a more accurate prediction of the variable of interest Y. The choice of q_i allows for some flexibility in how the regression coefficient is estimated.

To ensure the design-consistency of the projection estimator in (5), we can use the following regression estimator under non-nested two-phase sampling:

$$\widehat{Y}_{\mathrm{tp,reg}} = \widehat{Y}_2 + \left(\widehat{\mathbf{X}}_{\mathrm{c}} - \widehat{\mathbf{X}}_2 \right)' \hat{\boldsymbol{\beta}}_q \tag{6}$$

By the definition of $\widehat{\mathbf{X}}_{\mathrm{c}}$, we can also express this as:

$$\widehat{Y}_{\mathrm{tp,reg}} = \widehat{Y}_2 + \left(\widehat{\mathbf{X}}_1 - \widehat{\mathbf{X}}_2 \right)' \hat{\boldsymbol{\alpha}}_q, \tag{7}$$

where $\hat{\boldsymbol{\alpha}}_q = W \hat{\boldsymbol{\beta}}_q$. The key points are:

1. The design-consistent regression estimator $\widehat{Y}_{\text{tp,reg}}$ is constructed by adding a correction term to the projection estimator $\widehat{Y}_p$ from the second sample.

2. The regression estimator improves the efficiency of the design unbiased estimator $\hat{Y}_2$ by substracting the projection of $\hat{Y}_2$ onto the augmentation space (**?**), the linear space generated by the difference between the combined estimator $\widehat{\mathbf{X}}_{\text{c}}$ and the estimator $\widehat{\mathbf{X}}_2$ from the second sample.

3. Alternatively, the augmentation space can be expressed using the difference between the estimators $\widehat{\mathbf{X}}_1$ and $\widehat{\mathbf{X}}_2$, weighted by $\hat{\boldsymbol{\alpha}}_q$.

The goal is to leverage the information from both samples to obtain a design-consistent regression estimator for the variable of interest Y.

Using the standard argument, we can obtain

$$\widehat{Y}_{\text{tp,reg}} \;=\; \widehat{Y}_2 + \left(\widehat{\mathbf{X}}_1 - \widehat{\mathbf{X}}_2\right)' \boldsymbol{\alpha}_q^* + O_p(n^{-1}N) \tag{8}$$

where $\boldsymbol{\alpha}_q^*$ is the probability limit of $\hat{\boldsymbol{\alpha}}_q = W\hat{\boldsymbol{\beta}}_q$. By (8), we can obtain

$$V\left(\widehat{Y}_{\text{tp,reg}}\right) = (\boldsymbol{\alpha}_q^*)' V\left(\widehat{\mathbf{X}}_1\right) \boldsymbol{\alpha}_q^* + V\left(\hat{u}_2\right) \tag{9}$$

where $\hat{u}_2 = \sum_{i\in A_2} \pi_{2i}^{-1}\left(y_i - \mathbf{x}_i'\boldsymbol{\alpha}_q^*\right)$. From the formula in (9), we can construct a linearized variance estimator.

Now, we can use the calibration weighting to construct the regression estimator under non-nested two-phase sampling. For given the design weights $d_{2i} = \pi_{2i}^{-1}$, we find the minimizer of

$$Q\left(\boldsymbol{\omega}\right) = \sum_{i\in A_2} \left(\omega_i - d_{2i}\right)^2 q_i$$

subject to

$$\sum_{i\in A_2} \omega_i \mathbf{x}_i = \widehat{\mathbf{X}}_{\text{c}}.$$

The solution is

$$\hat{\omega}_i = d_{2i} + \left(\widehat{\mathbf{X}}_{\text{c}} - \widehat{\mathbf{X}}_2\right)^{-1} \left(\sum_{i\in A_2} q_i^{-1}\mathbf{x}_i\mathbf{x}_i'\right)^{-1} \mathbf{x}_i q_i^{-1}.$$

Note that

$$\sum_{i \in A_2} \hat{\omega}_i y_i = \widehat{Y}_{\text{tp,reg}},$$

where $\widehat{Y}_{\text{tp,reg}}$ is defined in (7). Thus, the algebraic equivalence between the regression estimator and the calibration weighting estimator is established under non-nested two-phase sampling.

Topic 3: Multi-source Two-Phase Sampling

References

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

Fuller, W. A. (2009). *Sampling statistics*. John Wiley & Sons.

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

Horvitz, D. G. and D. J. Thompson (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association 47*(260), 663–685.

Kim, J. K. (2024). *Statistics in Survey Sampling*. arXiv.

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

Narain, R. (1951). On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics 3*(2), 169–175.