Debiased Calibration for Generalized Two-Phase Sampling

Caleb Leedy

April 11, 2024

1 Introduction

- Combining information from several sources is an important practical problem.
- In many cases, we do not have direct access to the other sources. We can only obtain summary statistics (and their standard errors) for the external data sources. We wish to incorporate the information from external sources in our in-house data effectively using calibration weighting.
- We formulate the problem as a generalized two-phase sampling where the first phase sample can be obtained from multiple sources. The second-phase sample is our in-house data in which we want to construct the calibration weights.
- To achieve the goal, we first consider the classical two-phase sampling setup where the second-phase sample is a subset of the first-phase sample. After that, we extend the setup to more general cases such as non-nested two-phase sampling or multiple independent surveys with some common measurements.
- (To simplify the presentation, I wonder whether we can just use SRS in the first-phase sampling.)

2 Topic 1: Classical Two-Phase Sampling

2.1 Background and Introduction

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 auxilary 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 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,

$$\begin{aligned}\pi_i &= \sum_{A_2:i\in A_2} \Pr(A_2)\\ &= \sum_{A_1:A_2\subseteq A_1}\ \sum_{A_2:i\in A_2} \Pr(A_2 \mid A_1)\Pr(A_1)\\ &= \sum_{A_1:i\in A_1}\ \sum_{A_2:i\in A_2} \Pr(A_2 \mid A_1)\Pr(A_1).\end{aligned}$$

If we define $\pi_{2i|1}^{(2)} = \sum_{A_2:i\in A_2} \Pr(A_2 \mid A_1)$ and $\pi_i^{(1)} = \sum_{A_1:i\in A_1} \Pr(A_1)$ then,

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

(note: there is an abuse of notation. You do not need to introduce superscript (2) on $\pi_{2i|1}$.) This means that we can define the π^*-estimator as the following design unbiased estimator:

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

While unbiased Kim 2024, the π^*-estimator unfortunately 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 $\hat{X}_1$ obtained from the phase one sample. That is, we can leverage the external information $\hat{X}_1$ to improve the π^*-estimator in the second-phase sample. The two-phase regression estimator has the form,

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

where

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

I suggest that we use

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

where $q_i = q(\mathbf{x}_i)$ is a function of $\mathbf{x}_i$. This will have some connection with model-optimal regression estimator (under superpopulation model).

[1] Caleb, we do not have to use $\pi_{2i|1}^{(2)}$ in computing $\hat{\beta}_2$. Please check my book.

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

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

Using q_i, we can construct

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

as a way to implement the two-phase regression estimator indirectly using calibration weighting.

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. [2] This generalization includes empirical likelihood estimation, and maximum entropy estimation amoung 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, unless $\pi_{2i|1}^{-1} = x_i' a$ for some a, the estimator is not design consistent. [3]

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,

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

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, Kim, and Qiu 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.

Unfortunately, the method of Kwon, Kim, and Qiu 2024 requires known calibration levels (no uncertainty) for the finite population. It does not handle the two-phase setup where we

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

[3] 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.

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, Kim, and Qiu 2024 for the debiased calibration method. We consider maximizing the generalized entropy Gneiting and Raftery 2007,

$$H(w) = -\sum_{i \in A_2} G(w_i) \tag{1}$$

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

$$\sum_{i \in A_2} \frac{x_i w_i}{\pi_i^{(1)}} = \sum_{i \in A_1} \frac{x_i}{\pi_i^{(1)}} \tag{2}$$

and

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

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

Theoretical Results

We prove two results in this section is the same spirit as Kwon, Kim, and Qiu 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 $i = 1, \dots, n$,

[A3] If $\pi_{2ij|1}$ is joint inclusion probability of elements i and j and $\Delta_{2ij|1} = \pi_{2ij|1} - \pi_{2i|1}\pi_{2j|1}$ then,

$$\limsup_{n \to \infty} \max_{i,j \in A_1 : i \neq j} |\Delta_{2ij|1}| < \infty,$$

[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 $\lambda_0 = (\mathbf{0}, 1)$.

[A5] FIXME: We need to make an assumption about how Phase 2 relates to Phase 1 which relates to the finite population.

Theorem 1. *Design Consistency Suppose that Conditions [A1] - [A5] hold. Then the solution $\hat{w}$ to 1 subject to 2 and 3 exists and is unique with probability approaching 1. Furthermore, the estimator $\hat{\theta}_{DCE} = \sum_{i\in A_2} \frac{\hat{w} y_i}{\pi_i^{(1)}}$ satisfies*

$$\hat{\theta}_{DCE} = \hat{\theta}_{DC} + o_p(FIXME)$$

where $\hat{\theta}_{DC}$ is the debiased calibration estimator of Kwon, Kim, and Qiu 2024.

Theorem 2. *(Asymptotic Normality)*

Simulation Studies

Topic 2: Non-nested Two-Phase Sampling

Topic 3: Multi-source Two-Phase Sampling

References

Deville, Jean-Claude and Carl-Erik Sarndal (1992). “Calibration estimators in survey sampling”. In: *Journal of the American statistical Association* 87.418, pp. 376–382.

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

Gneiting, Tilmann and Adrian E Raftery (2007). “Strictly proper scoring rules, prediction, and estimation”. In: *Journal of the American statistical Association* 102.477, pp. 359–378.

Horvitz, Daniel G and Donovan J Thompson (1952). “A generalization of sampling without replacement from a finite universe”. In: *Journal of the American statistical Association* 47.260, pp. 663–685.

Kim, Jae Kwang (2024). *Statistics in Survey Sampling.* arXiv.

Kwon, Yonghyun, Jae Kwang Kim, and Yumou Qiu (2024). *Debiased calibration estimation using generalized entropy in survey sampling.* arXiv: `2404.01076 [stat.ME]`.

Narain, RD (1951). “On sampling without replacement with varying probabilities”. In: *Journal of the Indian Society of Agricultural Statistics* 3.2, pp. 169–175.