Debiased Calibration for Generalized Two-Phase Sampling

Caleb Leedy

May 17, 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 consists 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.

2 Topic 1: Classical Two-Phase Sampling

2.1 Background

Consider a finite population of size N containing elements $(\mathbf{X}_i, Y_i)$ where an initial (Phase 1) sample of size n_1 is selected and $\mathbf{X}_i$ is observed. Then from the Phase 1 sample of elements, a (Phase 2) sample of size $n_2 < n_1$ 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 $\bar{Y}_N$ that uses both the observed information in the Phase 2 sample and also the extra auxiliary information from $\mathbf{X}$ in the Phase 1 sample. 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,

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

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 $\mathbf{X}$. The two-phase regression estimator $\hat{Y}_{reg,tp}$ does incorporate information for $\mathbf{X}$ by using the estimate $\hat{\mathbf{X}}_1$ from the Phase 1 sample. This is how we can leverage the external information $\hat{\mathbf{X}}_{\mathbf{1}}$ to improve the initial π^*-estimator in the second phase sample. The two-phase regression estimator has the form,

$$\hat{Y}_{reg,tp} = \sum_{i\in A_1} \frac{1}{\pi_{1i}} \mathbf{x}_i \hat{\boldsymbol{\beta}}_q + \sum_{i\in A_2} \frac{1}{\pi_{1i}\pi_{2i|1}} (y_i - \mathbf{x}_i \hat{\boldsymbol{\beta}}_q)$$

where for $q_i = q(\mathbf{x}_i)$ and is a function of $\mathbf{x}_i$,

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

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

$$\hat{w}_{2i} = \arg\min_{w} \sum_{i \in A_2} (w_{2i} - \pi_{2i|1}^{-1})^2 q_i \text{ such that } \sum_{i \in A_2} w_{2i} \mathbf{x}_i / \pi_{1i} = \sum_{i \in A_1} \mathbf{x}_i / \pi_{1i}.$$

This means that $\hat{Y}_{reg,tp}$ is also a calibration estimator. The idea that regression estimation is a form of calibration was noted by Deville and Sarndal (1992) and extended by them to consider loss functions other than just squared loss. Their generalized loss function minimizes $\sum_i G(w_i, d_i) q_i$ for weights w_i and design-weights d_i where $G(\cdot)$ is a 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.[1] 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.

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

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

The Deville and Sarndal (1992) method incorporates the design weights into the loss function, which is the part minimizing the variance. We would rather separate have bias calibration separate from the minimizing the variance so that we can control each in isolation. 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 finite population calibration levels. It

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

does not handle the two-phase setup where we need to estimate the finite population total of $\mathbf{x}$ from the Phase 1 sample. In the rest of the section, we extend this method to two-phase sampling so that we have a valid estimator when including estimated Phase 1 weights with appropriate variance estimation.

2.2 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:

$$\sum_{i \in A_2} \frac{\mathbf{x}_i w_{2i} q_i}{\pi_{1i}} = \sum_{i \in A_1} \frac{\mathbf{x}_i q_i}{\pi_{1i}} \tag{2}$$

and

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

where $g(w) = \partial G / \partial w$.

The first constraint is the existing calibration constraint and the second ensures that design consistency is achieved. The original method of Kwon et al. (2024) only considered having finite population quantities on the right hand side of (2).

Writing $w_{1i} = \pi_{1i}^{-1}$, the the goal is to solve

$$\arg\min_{w_{2|1}} \sum_{i \in A_2} w_{1i} G(w_{2i}) q_i \text{ such that} \sum_{i \in A_2} w_{1i} w_{2i|1} \mathbf{z}_i q_i = \sum_{i \in A_1} w_{1i} \mathbf{z}_i q_i \tag{4}$$

where $\mathbf{z}_i = (\mathbf{x}_i / q_i, g(\pi_{2i|1}^{-1}))$. Let $\hat{w}_{2i|1}$ be the solution to Equation (4). The resulting estimator of Y_N is $\hat{Y}_{DCE} = \sum_{i \in A_2} w_{1i} \hat{w}_{2i|1} y_i$. To solve this problem we can use the method of Lagrange multipliers. We need to minimize the Lagrangian function

$$L(w_{2i|1}, \boldsymbol{\lambda}) = \sum_{i \in A_2} w_{1i} G(w_{2i|1}) q_i + \boldsymbol{\lambda} \left(\sum_{i \in A_1} w_{1i} \mathbf{z}_i q_i - \sum_{i \in A_2} w_{1i} w_{2i|1} \mathbf{z}_i q_i \right). \tag{5}$$

Differntiating with respect to $w_{2i|1}$ and setting this expression equal to zero, yields the fact that $\hat{w}_{2i|1}$ satisfies

$$\hat{w}_{2i|1}(\hat{\boldsymbol{\lambda}}) = g^{-1}(\hat{\boldsymbol{\lambda}}^T \mathbf{z}_i)$$

where $\hat{\boldsymbol{\lambda}}$ is the solution to

$$\left(\sum_{i \in A_1} w_{1i} \mathbf{z}_i q_i - \sum_{i \in A_2} w_{1i} w_{2i|1}(\hat{\boldsymbol{\lambda}}) \mathbf{z}_i q_i \right) = 0. \tag{6}$$

2.3 Theoretical Results

Theorem 1 (Design Consistency)**.** *Let $\boldsymbol{\lambda}^*$ be the probability limit of $\hat{\boldsymbol{\lambda}}$. Under some regularity conditions,*

$$\hat{Y}_{DCE} = \hat{Y}_{\ell}(\boldsymbol{\lambda}^*, \boldsymbol{\phi}^*) + O_p(N/n_2)$$

where

$$\hat{Y}_{\ell}(\boldsymbol{\lambda}, \boldsymbol{\phi}^*) = \hat{Y}_{DCE}(\hat{\boldsymbol{\lambda}}) + \left(\sum_{i \in A_1} w_{1i} \mathbf{z}_i q_i - \sum_{i \in A_2} w_{1i} \hat{w}_{2i|1}(\boldsymbol{\lambda}) z_i q_i \right) \boldsymbol{\phi}^*$$

and

$$\boldsymbol{\phi}^* = \left[\sum_{i \in U} \frac{\pi_{2i|1} \mathbf{z}_i \mathbf{z}_i^T q_i}{g'(d_{2i|1})} \right]^{-1} \sum_{i \in U} \frac{\pi_{2i|1} \mathbf{z}_i y_i}{g'(d_{2i|1})}.$$

$$\hat{Y}_{\ell}(\boldsymbol{\lambda}^*, \boldsymbol{\phi}^*) \;\; = \;\; \hat{Y}_{\pi^*} + \left(\sum_{i \in A_1} w_{1i} \mathbf{x}_i - \sum_{i \in A_2} w_{1i} \pi_{2i|1}^{-1} \mathbf{x}_i \right)^T \boldsymbol{\phi}_1^* + \left(\sum_{i \in A_1} w_{1i} g_i - \sum_{i \in A_2} w_{1i} \pi_{2i|1}^{-1} g_i \right)^T \phi_2^*$$

and

$$\begin{pmatrix}\boldsymbol{\phi}_1^* \\ \phi_2^*\end{pmatrix} = \left[\sum_{i\in U}\frac{\pi_{2i|1}}{g'(d_{2i|1})q_i}\begin{pmatrix}\mathbf{x}_i\mathbf{x}_i^T & \mathbf{x}_i g_i \\ g_i\mathbf{x}_i^T & g_i^2\end{pmatrix}\right]^{-1}\sum_{i\in U}\frac{\pi_{2i|1}}{g'(d_{2i|1})q_i}\begin{pmatrix}\mathbf{x}_i \\ g_i\end{pmatrix}y_i$$

with $g_i = g(\pi_{2i|1}^{-1})q_i$*.*

Proof. In this proof, we derive the solution to Equation 4 and show that it is asymptotically equivalent to a regression estimator. Using the method of Lagrange multipliers, to solve Equation (4) we need to minimize the Lagrangian in Equation (5). The first order conditions show that

$$\frac{\partial\mathcal{L}}{\partial w_{2i|1}} : g(w_{2i|1})w_{1i}q_i - w_{1i}\boldsymbol{\lambda}\mathbf{z}_i q_i = 0.$$

Hence, $\hat{w}_{2i}(\boldsymbol{\lambda}) = g^{-1}(\boldsymbol{\lambda}^T\mathbf{z}_i)$ and $\hat{\boldsymbol{\lambda}}$ is determined by Equation (6). When the sample size gets large, we have $\hat{w}_{2i|1}(\hat{\boldsymbol{\lambda}}) \to d_{2i|1}$ which means that $\hat{\boldsymbol{\lambda}} \to \boldsymbol{\lambda}^*$ where $\boldsymbol{\lambda}^* = (\mathbf{0}, 1)$. Then using the linearization technique of Randles (1982), we can construct a regression estimator,

$$\hat{Y}_\ell(\hat{\boldsymbol{\lambda}}, \boldsymbol{\phi}) = \hat{Y}_{DCE}(\hat{\boldsymbol{\lambda}}) + \left(\sum_{i\in A_1} w_{1i}\boldsymbol{z}_i q_i - \sum_{i\in A_2} w_{1i}\hat{w}_{2i|1}(\hat{\boldsymbol{\lambda}})\mathbf{z}_i q_i\right)\boldsymbol{\phi}.$$

Notice that $\hat{Y}_\ell(\hat{\boldsymbol{\lambda}}, \boldsymbol{\phi}) = \hat{Y}_{DCE}(\hat{\boldsymbol{\lambda}})$ for all $\boldsymbol{\phi}$. We choose $\boldsymbol{\phi}^*$ such that

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

Using the fact that $g^{-1}(\boldsymbol{\lambda}^*\mathbf{z}_i) = g^{-1}(g(d_{2i|1})) = d_{2i|1}$ and $(g^{-1})'(x) = 1/g'(g^{-1}(x))$, we have

$$\begin{aligned}\boldsymbol{\phi}^* &= E\left[\sum_{i\in A_2}\frac{w_{1i}\mathbf{z}_i\mathbf{z}_i^T q_i}{g'(d_{2i|1})}\right]^{-1} E\left[\sum_{i\in A_2}\frac{w_{1i}\mathbf{z}_i y_i}{g'(d_{2i|1})}\right] \\ &= \left[\sum_{i\in U}\frac{\pi_{2i|1}\mathbf{z}_i\mathbf{z}_i^T q_i}{g'(d_{2i|1})}\right]^{-1}\left[\sum_{i\in U}\frac{\pi_{2i|1}\mathbf{z}_i y_i}{g'(d_{2i|1})}\right]\end{aligned}$$

Thus, the linearization estimator is

$$\hat{Y}_\ell(\boldsymbol{\lambda}^*, \boldsymbol{\phi}^*) = \sum_{i \in A_1} w_{1i} q_i \mathbf{z}_i \boldsymbol{\phi}^* + \sum_{i \in A_2} w_{1i} d_{2i|1} (y_i - q_i \mathbf{z}_i \boldsymbol{\phi}^*).$$

By construction using a Taylor expansion yields,

$$\begin{aligned} \hat{Y}_{DCE}(\hat{\boldsymbol{\lambda}}) &= \hat{Y}_\ell(\boldsymbol{\lambda}^*, \boldsymbol{\phi}^*) + E\left[\frac{\partial}{\partial \boldsymbol{\lambda}} \hat{Y}_\ell(\boldsymbol{\lambda}^*, \boldsymbol{\phi}^*)\right](\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^*) + \frac{1}{2} E\left[\frac{\partial}{\partial \boldsymbol{\lambda}^2} \hat{Y}_{DCE}(\boldsymbol{\lambda}^*)\right](\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^*)^2 \\ &= \hat{Y}_\ell(\boldsymbol{\lambda}^*, \boldsymbol{\phi}^*) + O(N) O_p(n_2^{-1}). \end{aligned}$$

The final equality comes from the fact that $E\left[\frac{\partial}{\partial \boldsymbol{\lambda}} \hat{Y}_\ell(\boldsymbol{\lambda}^*, \boldsymbol{\phi}^*)\right] = 0$, $\frac{\partial}{\partial \boldsymbol{\lambda}^2} \hat{w}_{2i|1}(\boldsymbol{\lambda}^*)$ is bounded and $||\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^*|| = O_p(n_2^{-1/2})$, which proves our result. □

2.4 Simulation Study 1

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

$$\begin{aligned} X_{1i} &\overset{ind}{\sim} N(2, 1) \\ X_{2i} &\overset{ind}{\sim} Unif(0, 4) \\ X_{3i} &\overset{ind}{\sim} N(0, 1) \\ \varepsilon_i &\overset{ind}{\sim} N(0, 1) \\ Y_i &= 3X_{1i} + 2X_{2i} + \varepsilon_i \\ \pi_{1i} &= n_1/N \\ \pi_{2i|1} &= \min(\Phi_3(-x_{1i} + 1), 0.9). \end{aligned}$$

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 the Phase 1 sampling being a simple random sample (SRS) and the Phase 2 sampling occuring under Poisson sampling. This yields a Phase 1 sample size of $n_1 = 1000$ and a Phase 2 sample size of $E[n_2] \approx 267$. 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 $\mathbf{x}_i = (1, x_{1i}, x_{2i})^T$,

3. Debiased Calibration with Population Constraints (DC-Pop): This solves

$$\underset{w_{2|1}}{\arg\min} \sum_{i \in A_2} w_{1i} G(w_{2i}) \text{ such that} \sum_{i \in A_2} w_{1i} w_{2i|1} \mathbf{z}_i = \sum_{i \in U} \mathbf{z}_i. \tag{7}$$

4. Debiased Calibration with Estimated Population Constraints (DC-Est): This solves Equation (4) with $q_i = 1$.

5. Linearized Debiased Calibration with Population Constraints (Eta-Pop): This is the linearized estimate DC-Pop using the linearization in Theorem 1.

6. Linearized Debiased Calibration with Estimated Population Constraints (Eta-Est): This is the linearized estimate DC-Est using the linearization in Theorem 1.

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 give the variance estimate in Table 1.

Estimator	Estimated Variance	Notes
π^*	$\left(\frac{1}{n_1} - \frac{1}{N}\right) \hat{s}_Y^2 + \sum_{i \in A_2} \frac{w_{1i}^2}{N^2} \frac{(1-\pi_{2i\|1})}{\pi_{2i\|1}^2} y_i^2$	$\hat{s}_Y^2 = \frac{1}{n_2-1} \sum_{i \in A_2} (y_i - \bar{y}_i)^2$
TP-Reg	$\left(\frac{1}{n_1} - \frac{1}{N}\right) \hat{s}_\eta^2 + \sum_{i \in A_2} \frac{w_{1i}}{N^2} \frac{(1-\pi_{2i\|1})}{\pi_{2i\|1}^2} \varepsilon_i^2$	$\eta_i = \mathbf{x}_i' \boldsymbol{\beta} + \delta_{2i} w_{2i\|1} \varepsilon_i, \ \varepsilon_i = (y_i - \mathbf{x}_i' \boldsymbol{\beta})$
DC-Pop	$\left(\frac{1}{n_1} - \frac{1}{N}\right) \hat{s}_\varepsilon^2 + \sum_{i \in A_2} \frac{w_{1i}}{N^2} \frac{(1-\pi_{2i\|1})}{\pi_{2i\|1}^2} \varepsilon_i^2$	$\eta_i = \mathbf{z}_i' \boldsymbol{\phi} + \delta_{2i} w_{2i\|1} \varepsilon_i, \ \varepsilon_i = (y_i - \mathbf{z}_i' \boldsymbol{\phi})$
DC-Est	$\left(\frac{1}{n_1} - \frac{1}{N}\right) \hat{s}_\eta^2 + \sum_{i \in A_2} \frac{w_{1i}}{N^2} \frac{(1-\pi_{2i\|1})}{\pi_{2i\|1}^2} \varepsilon_i^2$	$\eta_i = \mathbf{z}_i' \boldsymbol{\phi} + \delta_{2i} w_{2i\|1} \varepsilon_i, \ \varepsilon_i = (y_i - \mathbf{z}_i' \boldsymbol{\phi})$
Eta-Pop	Same as DC-Pop	
Eta-Est	Same as DC-Est	

Table 1: This table gives the formulas for each variance estimator used in Simulation 1. For the derivation of the estimated variance of the DC-Pop estimatator see Appendix A.

We run the simulation 1000 times for each of these methods and compute the Bias ($E[\hat{Y}] - \bar{Y}_N$), the RMSE ($\sqrt{\mathrm{Var}(\hat{Y} - \bar{Y}_N)}$), 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 where $\hat{Y}^{(b)}$ is the result from the bth simulation replicate. The results are in Table 2.

Est	Bias	RMSE	EmpCI	Ttest
π^*	-0.064	1.045	0.928	1.930
TP-Reg	-0.003	0.131	0.945	0.755
DC-Pop	0.001	0.102	0.962	0.340
DC-Est	-0.002	0.146	0.942	0.349
Eta-Pop	0.001	0.102	0.962	0.340
Eta-Est	-0.001	0.146	0.942	0.193

Table 2: This table shows the results of Simulation Study 1. 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.

While DC-Est still does not outperform TP-Reg, it is equivalent to Eta-Est, which is a regression estimator.

The variance estimates look good now. I needed to think of DC-Pop as a regression estimator instead of a two-phase regression estimator.

I also give the mean variance estimates, the Monte Carlo mean, as well as a test for bias.

Est	MCVar	EstVar	VarVar	Ttest
π^*	1.0900	1.1116	0.2184	1.4571
TP-Reg	0.0170	0.0174	0.0000	3.9638
DC-Pop	0.0104	0.0109	0.0006	0.5924
DC-Est	0.0213	0.0195	0.0011	1.7745
Eta-Pop	0.0104	0.0109	0.0006	0.5924
Eta-Est	0.0214	0.0195	0.0011	1.9081

Table 3: This table shows the variance estimates from Simulation Study 1. It displays the Monte Carlo variance of $\hat{Y}^{(b)}$, the average estimated variance, the variance of the variance estimator, and a t-statistic assessing the unbiasedness of each variance estimator.

3 Topic 2: Non-nested Two-Phase Sampling

3.1 Background

Now we consider the sampling mechanism known as non-nested two-phase sampling (Hidiroglou (2001)). In the last section, we considered two-phase sampling in which the Phase 2 sample was a subset of the Phase 1 sample. With non-nested two-phase sampling the Phase 2 sample is independent of the Phase 1 sample. It is a separate independent sample of the same population. Like traditional two phase sampling, we consider the Phase 1 sample, A_1, to consist of observations of $(\mathbf{X}_i)_{i=1}^{n_1}$ and the Phase 2 sample, A_2, to consist of observations of $(\mathbf{X}_i, Y_i)_{i=1}^{n_2}$.

Whereas the classical two-phase estimator uses a single Horvitz-Thompson estimator of the Phase 1 sample to construct estimates for calibration totals, in the non-nested two-phase sample we have two independent Horvitz-Thompson estimators of the total of $\mathbf{X}$,

$$\hat{\mathbf{X}}_1 = \sum_{i \in A_1}^{n_1} d_{1i}\mathbf{x}_i \text{ and } \hat{\mathbf{X}}_2 = \sum_{i \in A_2}^{n_2} d_{2i}\mathbf{x}_i$$

where $d_{1i} = \pi_{1i}^{-1}$, $d_{2i} = \pi_{2i}^{-1}$, π_{1i} is the probability of $i \in A_1$ and $\pi_{2i} = \Pr(i \in A_2)$. We can combine these estimates using the effective sample size (Kish (1965)) to get $\hat{\mathbf{X}}_c = (n_{1,eff}\hat{\mathbf{X}}_1 + n_{2,eff}\hat{\mathbf{X}}_2)/(n_{1,eff} + n_{2,eff})$ where $n_{1,eff}$ and $n_{2,eff}$ are the effective sample size for A_1 and A_2 respectively. Then we can define a regression estimator as

$$\hat{Y}_{NN,reg} = \hat{Y}_2 + (\hat{\mathbf{X}}_c - \hat{\mathbf{X}}_2)^T\hat{\boldsymbol{\beta}}_q = \hat{Y}_2 + (\hat{\mathbf{X}}_1 - \hat{\mathbf{X}}_2)^T W\hat{\boldsymbol{\beta}}_q$$

where, $W = n_{1,eff}/(n_{1,eff} + n_{2,eff})$, and

$$\hat{\boldsymbol{\beta}}_q = \left(\sum_{i \in A_2} \frac{\mathbf{x}_i\mathbf{x}_i^T}{q_i}\right)^{-1} \sum_{i \in A_2} \frac{\mathbf{x}_i y_i}{q_i} \text{ and } \hat{Y}_2 = \sum_{i \in A_2} d_{2i}y_i.$$

Since the samples A_1 and A_2 are independent,

$$V(\hat{Y}_{NN,reg}) = V\left(\sum_{i \in A_2} \frac{1}{\pi_{2i}}(y_i - \mathbf{x}_i^T W\boldsymbol{\beta}_q^*)\right) + (\boldsymbol{\beta}_q^*)^T W^T V(\hat{\mathbf{X}}_1) W\boldsymbol{\beta}_q^*$$

where $\boldsymbol{\beta}_q^*$ is the probability limit of $\hat{\boldsymbol{\beta}}_q$. Like the two-phase sample this regression estimator can be viewed as the solution to the following calibration equation

$$\hat{w}_2 = \arg\min_{w_2} Q(w_2) = \sum_{i \in A_2} (w_{2i} - d_{2i})^2 q_i \text{ such that } \sum_{i \in A_2} w_{2i} \mathbf{x}_i = \hat{\mathbf{X}}_c \tag{8}$$

and $\hat{Y}_{NN,reg} = \sum_{i \in A_2} \hat{w}_{2i} y_i$ where $\hat{w}_{2i}$ is the solution to Equation 8.

We extend the debiased calibration estimator of Kwon et al. (2024) to the non-nested two-phase sampling case where we use a combined estimate $\hat{\boldsymbol{X}}_c$ as the calibration totals instead of using the true totals from the finite population.

3.2 Methodology

The methodology for the non-nested two-phase sample is very similar to the setup described as part of Topic 1. Given a strictly convex differentiable function, $G : \mathcal{V} \to \mathbb{R}$, the goal is to solve

$$\hat{w}_2 = \arg\min_{w} \sum_{i \in A_2} G\left(w_{2i}\right) q_i \text{ with } \sum_{i \in A_2} w_{2i} \mathbf{x}_i = \hat{\mathbf{X}}_c \text{ and } \sum_{i \in A_2} w_{2i} g(d_{2i}) q_i = \sum_{i \in U} g(d_{2i}) q_i \tag{9}$$

for $g(x) = G'(x)$ and a known choice of $q_i \in \mathbb{R}$. The difference between solving Equation 9 and Equation 4 is that the estimator $\hat{\mathbf{X}}_c$ is estimated from the combined sample $A_c = A_1 \cup A_2$. Before using $\hat{\mathbf{X}}_c$ in the debiased calibration estimator, we need to estimate it from the non-nested samples. We can get multiple estimates of $\hat{\mathbf{X}}$,

$$\hat{\mathbf{X}}_1 = \sum_{i \in A_1} d_{1i} \mathbf{x}_i \text{ and } \hat{\mathbf{X}}_2 = \sum_{i \in A_2} d_{2i} \mathbf{x}_i.$$

Let V_1 and V_2 be known variance matrices for $\hat{\mathbf{X}}_1$ and $\hat{\mathbf{X}}_2$, then the optimal combined estimate is

$$\hat{\mathbf{X}}_c = \frac{V_1^{-1} \hat{\mathbf{X}}_1 + V_2^{-1} \hat{\mathbf{X}}_2}{V_1^{-1} + V_2^{-1}}.$$

We can constuct a non-nested two-phase estimator $\hat{Y}_{NNE}$ for Y_N where $\hat{Y}_{NNE} = \sum_{i \in A_2} \hat{w}_{2i} y_i$ and $\hat{w}_{2i}$ solves Equation 9. Like the classical two-phase approach, to solve this setup we minimize the Lagrangian,

$$L(w_{2i}, \boldsymbol{\lambda}) = \sum_{i \in A_2} G(w_{2i}) q_i + \boldsymbol{\lambda} \left(\hat{\mathbf{T}} - \sum_{i \in A_2} w_{2i} \mathbf{z}_i q_i \right). \tag{10}$$

with

$$\hat{\mathbf{T}} = \begin{bmatrix} \hat{\mathbf{X}}_c \\ \sum_{i \in U} g(d_{2i}) q_i \end{bmatrix}.$$

I am a little worried about the second term of $\hat{T}$. It is somewhat weird to assume that $T_g = \sum_{i \in U} g(d_{2i}) q_i$ is known. One simple way is to use a HT estimator of T_g from sample A_2. That is, use $\hat{T}_{g,\mathrm{HT}} = \sum_{i \in A_2} d_{2i} g(d_{2i}) q_i$ or even use a GREG estimator of T_g.

Differntiating with respect to w_{2i} and setting this expression equal to zero, yields the fact that $\hat{w}_{2i}$ satisfies

$$\hat{w}_{2i}(\hat{\boldsymbol{\lambda}}) = g^{-1}(\hat{\boldsymbol{\lambda}}^T \mathbf{z}_i)$$

where $\hat{\boldsymbol{\lambda}}$ is the solution to

$$\left(\hat{\mathbf{T}} - \sum_{i \in A_2} w_{2i}(\hat{\boldsymbol{\lambda}}) \mathbf{z}_i q_i \right) = 0. \tag{11}$$

3.3 Theoretical Results

Theorem 2 (Design Consistency). *Allowing $\boldsymbol{\lambda}^*$ to be the probability limit of $\hat{\boldsymbol{\lambda}}$, under some regularity conditions, $\hat{Y}_{NNE} = \hat{Y}_{\ell,NNE}(\boldsymbol{\lambda}^*, \boldsymbol{\phi}^*) + O_p(N n_2^{-1})$ where*

$$\hat{Y}_{\ell,NNE}(\boldsymbol{\lambda}^*, \boldsymbol{\phi}^*) = \sum_{i \in A_2} \hat{w}_{2i}(\boldsymbol{\lambda}^*) + \left(\hat{\mathbf{T}} - \sum_{i \in A_2} \hat{w}_{2i}(\boldsymbol{\lambda}^*) \mathbf{z}_i q_i \right) \boldsymbol{\phi}^*$$

and

$$\boldsymbol{\phi}^* = \left(\sum_{i \in U} \frac{\pi_{2i} q_i}{g'(d_{2i})} \begin{bmatrix} \mathbf{x}_i^2/q_i & \mathbf{x}_i g(d_{2i})/q_i \\ \mathbf{x}_i g(d_{2i})/q_i & g(d_{2i})^2 \end{bmatrix} \right)^{-1} \sum_{i \in U} \frac{\pi_{2i} y_i}{g'(d_{2i})} \begin{bmatrix} \mathbf{x}_i/q_i \\ g(d_i) \end{bmatrix}.$$

Proof. The proof of this result is very similar to the proof the Theorem 1. The biggest difference is that the total for $\mathbf{X}$ is estimated from both samples using $\hat{\mathbf{X}}_c$ instead of $\hat{\mathbf{X}}_{HT}$ from the Phase 1 sample.

Since $\hat{Y}_{NNE} = \sum_{i \in A_2} \hat{w}_{2i}(\hat{\boldsymbol{\lambda}}) y_i$ where $\hat{\boldsymbol{\lambda}}$ solves

$$\sum_{i \in A_2} \hat{w}_{2i}(\boldsymbol{\lambda}) q_i \underbrace{\begin{bmatrix} \mathbf{x}_i/q_i \\ g(d_i) \end{bmatrix}}_{\mathbf{z}_i} = \mathbf{T} \tag{12}$$

we have

$$\hat{Y}_{\ell,NNE}(\hat{\boldsymbol{\lambda}}, \boldsymbol{\phi}) = \sum_{i \in A_2} \hat{w}_{2i}(\hat{\boldsymbol{\lambda}}) + \left(\mathbf{T} - \sum_{i \in A_2} \hat{w}_{2i}(\hat{\boldsymbol{\lambda}}) \mathbf{z}_i q_i \right) \boldsymbol{\phi}.$$

If we choose $\boldsymbol{\phi}^*$ such that $E\left[\frac{\partial}{\partial \boldsymbol{\lambda}} \hat{Y}_{\ell,NNE}(\boldsymbol{\lambda}^*, \boldsymbol{\phi}^*)\right] = 0$, then

$$\boldsymbol{\phi}^* = \begin{bmatrix} \boldsymbol{\phi}_1^* \\ \boldsymbol{\phi}_2^* \end{bmatrix} = \left(\sum_{i \in U} \frac{\pi_{2i} q_i}{g'(d_{2i})} \begin{bmatrix} \mathbf{x}_i^2/q_i & \mathbf{x}_i g(d_{2i})/q_i \\ \mathbf{x}_i g(d_{2i})/q_i & g(d_{2i})^2 \end{bmatrix} \right)^{-1} \sum_{i \in U} \frac{\pi_{2i} y_i}{g'(d_{2i})} \begin{bmatrix} \mathbf{x}_i/q_i \\ g(d_i) \end{bmatrix}.$$

Hence, by a Taylor expansion around $\hat{\boldsymbol{\lambda}}$,

$$\hat{Y}_{NNE}(\hat{\boldsymbol{\lambda}}) = \hat{Y}_{\ell,NNE}(\boldsymbol{\lambda}^*, \boldsymbol{\phi}^*) + O_p(N n_2^{-1}).$$

□

Theorem 3 (Variance Estimation)**.** *The variance of* $\hat{Y}_{NNE}$ *is*

$$\begin{aligned} Var(\hat{Y}_{NNE}(\hat{\lambda}) = (\boldsymbol{\phi}_1^*)^T \, Var(\hat{\mathbf{X}}_c) \boldsymbol{\phi}_1^* + \sum_{i \in U} \sum_{j \in U} \frac{\Delta_{2ij}}{\pi_{2i} \pi_{2j}} (y_i - \mathbf{z}_i \boldsymbol{\phi}^* q_i)(y_j - \mathbf{z}_j \boldsymbol{\phi}^* q_j) \\ + (1 - W) \boldsymbol{\phi}_1^* \sum_{i \in U} \sum_{j \in U} \Delta_{2ij} d_{2i} \mathbf{x}_i d_{2j} (y_j - \mathbf{z}_j \boldsymbol{\phi}_1^* q_j) \end{aligned}$$

We can estimate the variance using

$$\hat{V}_{NNE} = (\hat{\boldsymbol{\phi}}_1)^T \, Var(\hat{\mathbf{X}}_c)\hat{\boldsymbol{\phi}}_1 + \sum_{i\in A_2}\sum_{j\in A_2}\frac{\Delta_{2ij}}{\pi_{2ij}\pi_{2i}\pi_{2j}}(y_i - \mathbf{z}_i\hat{\boldsymbol{\phi}}q_i)(y_j - \mathbf{z}_j\hat{\boldsymbol{\phi}}q_j)$$
$$+ (1-W)\hat{\boldsymbol{\phi}}_1\sum_{i\in A_2}\sum_{j\in A_2}\frac{\Delta_{2ij}}{\pi_{2ij}}\frac{\mathbf{x}_i}{\pi_{2i}}\frac{(y_j - \mathbf{z}_j\hat{\boldsymbol{\phi}}_1 q_j)}{\pi_{2j}}$$

where

$$\hat{\boldsymbol{\phi}} = \begin{bmatrix}\hat{\boldsymbol{\phi}}_1\\ \hat{\phi}_2\end{bmatrix} = \left(\sum_{i\in A_2}\frac{q_i}{g'(d_{2i})}\begin{bmatrix}\mathbf{x}_i^2/q_i & \mathbf{x}_i g(d_{2i})/q_i\\ \mathbf{x}_i g(d_{2i})/q_i & g(d_{2i})^2\end{bmatrix}\right)^{-1}\sum_{i\in A_2}\frac{y_i}{g'(d_{2i})}\begin{bmatrix}\mathbf{x}_i/q_i\\ g(d_i)\end{bmatrix}.$$

Proof. From Theorem 2, we know that $\hat{Y}_{NNE}(\hat{\lambda}) = \hat{Y}_{\ell,NNE}(\boldsymbol{\lambda}^*, \boldsymbol{\phi}^*) + O_p(Nn_2^{-1})$. Hence, the variance of $\hat{Y}_{NNE}(\hat{\boldsymbol{\lambda}})$ is

$$\begin{aligned}
\mathrm{Var}(\hat{Y}_{NNE}(\hat{\boldsymbol{\lambda}})) &= \mathrm{Var}(\hat{Y}_{\ell,NNE}(\boldsymbol{\lambda}^*, \boldsymbol{\phi}^*) + O_p(Nn_2^{-1}))\\
&= \mathrm{Var}\left(\sum_{i\in A_2}\hat{w}_{2i}(\boldsymbol{\lambda}^*)y_i + \left(\mathbf{T} - \sum_{i\in A_2}\hat{w}_{2i}(\boldsymbol{\lambda}^*)\mathbf{z}_i q_i\right)\boldsymbol{\phi}^*\right)\\
&= (\boldsymbol{\phi}_1^*)^T\mathrm{Var}(\hat{\mathbf{X}}_c)\boldsymbol{\phi}_1^* + \sum_{i\in U}\sum_{j\in U}\frac{\Delta_{2ij}}{\pi_{2i}\pi_{2j}}(y_i - \mathbf{z}_i\boldsymbol{\phi}^* q_i)(y_j - \mathbf{z}_j\boldsymbol{\phi}^* q_j)\\
&\quad + 2\mathrm{Cov}\left(\hat{\mathbf{X}}_c\boldsymbol{\phi}_1^*, \sum_{i\in A_2}\frac{(y_i - \mathbf{z}_i\boldsymbol{\phi}^* q_i)}{\pi_{2i}}\right)\\
&= (\boldsymbol{\phi}_1^*)^T\mathrm{Var}(\hat{\mathbf{X}}_c)\boldsymbol{\phi}_1^* + \sum_{i\in U}\sum_{j\in U}\frac{\Delta_{2ij}}{\pi_{2i}\pi_{2j}}(y_i - \mathbf{z}_i\boldsymbol{\phi}^* q_i)(y_j - \mathbf{z}_j\boldsymbol{\phi}^* q_j)\\
&\quad + (1-W)\boldsymbol{\phi}_1^*\sum_{i\in U}\sum_{j\in U}\Delta_{2ij}\frac{x_i}{\pi_{2i}}\frac{(y_j - \mathbf{z}_j\boldsymbol{\phi}_1^* q_j)}{\pi_{2j}}
\end{aligned}$$

where the last equality comes from the fact that $\hat{\mathbf{X}}_c = W\hat{\mathbf{X}}_1 + (1-W)\hat{\mathbf{X}}_2$. To have an unbiased estimator of the variance we can use:

$$
\begin{aligned}
\hat{V}_{NNE} = (\hat{\boldsymbol{\phi}}_1)^T \mathrm{Var}(\hat{\mathbf{X}}_c)\hat{\boldsymbol{\phi}}_1 + \sum_{i \in A_2}\sum_{j \in A_2} \frac{\Delta_{2ij}}{\pi_{2ij}\pi_{2i}\pi_{2j}}(y_i - \mathbf{z}_i\hat{\boldsymbol{\phi}}q_i)(y_j - \mathbf{z}_j\hat{\boldsymbol{\phi}}q_j) \\
+ (1-W)\hat{\phi}_1 \sum_{i \in A_2}\sum_{j \in A_2} \frac{\Delta_{2ij}}{\pi_{2ij}}\frac{x_i}{\pi_{2i}}\frac{(y_j - \mathbf{z}_j\hat{\boldsymbol{\phi}}_1 q_j)}{\pi_{2j}}
\end{aligned}
$$

where

$$
\hat{\boldsymbol{\phi}} = \begin{bmatrix} \hat{\boldsymbol{\phi}}_1 \\ \hat{\phi}_2 \end{bmatrix} = \left(\sum_{i \in A_2} \frac{q_i}{g'(d_{2i})} \begin{bmatrix} \mathbf{x}_i^2/q_i & \mathbf{x}_i g(d_{2i})/q_i \\ \mathbf{x}_i g(d_{2i})/q_i & g(d_{2i})^2 \end{bmatrix} \right)^{-1} \sum_{i \in A_2} \frac{y_i}{g'(d_{2i})} \begin{bmatrix} \mathbf{x}_i/q_i \\ g(d_i) \end{bmatrix}.
$$

□

3.4 Simulation Study 2

We run a simulation testing the proposed method. This is very similar to Simulation 1. We have the following simulation setup:

$$
\begin{aligned}
X_{1i} &\overset{ind}{\sim} N(2,1) \\
X_{2i} &\overset{ind}{\sim} Unif(0,4) \\
X_{3i} &\overset{ind}{\sim} N(0,1) \\
\varepsilon_i &\overset{ind}{\sim} N(0,1) \\
Y_i &= 3X_{1i} + 2X_{2i} + \varepsilon_i \\
\pi_{1i} &= n_1/N \\
\pi_{2i} &= \min(\Phi_3(-x_{1i}+1), 0.9).
\end{aligned}
$$

where Φ_3 is the CDF of a t-distribution with 3 degrees of freedom. We consider a finite population of size $N = 10,000$ with the Phase 1 sampling being a simple random sample (SRS) of size $n_1 = 1000$. The Phase 2 sample is a Poisson sample with an expected sample size of

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 $E[n_2] \approx 300$. 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. HT-estimator: $\hat{Y}_{HT} = N^{-1} \sum_{i \in A_2} \frac{y_i}{\pi_{2i}}$,

2. Regression estimator (Reg): Let $\hat{Y}_{NN,reg} = \hat{Y}_{HT} + (\hat{\mathbf{X}}_c - \hat{\mathbf{X}_{2,HT}})\hat{\boldsymbol{\beta}}_2$ where $\hat{Y}_{HT} = \sum_{i \in A_2} d_{2i} y_i$, $\hat{\mathbf{X}}_c = W\hat{\mathbf{X}_{1,HT}} + (1 - W)\hat{\mathbf{X}_{2,HT}}$, $W = n_{1,eff}/(n_{1,eff} + n_{2,eff})$, $\hat{\mathbf{X}_{1,HT}} = \sum_{i \in A_1} d_{1i}\mathbf{x}_i$, $\hat{\mathbf{X}_{2,HT}} = \sum_{i \in A_2} d_{2i}\mathbf{x}_i$, $\mathbf{x}_i = (1, x_{1i}, x_{2i})^T$ and

$$\hat{\boldsymbol{\beta}}_2 = \left(\sum_{i \in A_2} \mathbf{x}_i \mathbf{x}_i^T\right)^{-1} \sum_{i \in A_2} \mathbf{x}_i y_i.$$

   Then $\hat{\bar{Y}}_{NN,reg} = \hat{Y}_{NN,reg}/N$.

3. Debiased Calibration with Population Constraints (DC-Pop): This solves

$$\hat{w}_2 = \arg\min_{w} \sum_{i \in A_2} G(w_{2i})q_i \text{ with } \sum_{i \in A_2} w_{2i}\mathbf{x}_i = \sum_{i \in U} \mathbf{x}_i \text{ and } \sum_{i \in A_2} w_{2i} g(d_{2i}) q_i = \sum_{i \in U} g(d_{2i}) q_i$$

4. Debiased Calibration with Estimated Population Constraints (DC-Est): This solves Equation (9) with $q_i = 1$.

In addition to estimating the mean parameter $\bar{Y}_N$, we also construct variance estimates $\hat{V}(\hat{Y})$ for each estimate $\hat{Y}$.

We run the 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 where $\hat{Y}^{(b)}$ is the result from the bth simulation replicate. The results are in Table 4.

Est	Bias	RMSE	EmpCI	Ttest
HT	-0.007	0.372	0.949	0.630
Reg	-0.004	0.111	0.957	1.233
DC-Pop	0.000	0.028	0.939	0.031
DC-Est	-0.004	0.111	0.955	1.252

Table 4: This table shows the results of Simulation Study 2. It displays the Bias, RMSE, empirical 95% confidence interval, and a t-statistic assessing the unbiasedness of each estimator for the estimators: HT, Reg, DC-Pop, and DC-Est.

4 Topic 3: Multi-Source Two-Phase Sampling

4.1 Background

When considering non-nested two-phase sampling, we focused on the case of having two samples. Now, we consider incorporating more than two independent samples together with a debiasing constraint. First, we consider the case in which we want to estimate the $\theta = E[Y]$ and Y is only observed in one sample. Then we will consider the case where Y may be in multiple samples.

Consider the setup in which we have independent samples $A_0, A_1, \dots, A_M$ where Y is observed only in A_0 but $\mathbf{X}$ is observed in all of the samples with elements $\mathbf{X}^{(0)}$ observed in A_0 and $\mathbf{X}^{(m)}$ observed in A_m for each $m = 1, \dots, M$. Like the non-nested case, we assume that each survey is sampling independently from the same population sampling frame.

The traditional multi-source approach (Kim (2024)) is to use generalized least squares (GLS) to obtain an optimal estimator of X from the samples $A_1, \dots, A_M$. Then we can incorporate this information in the estimation of θ by using the following estimate

$$\hat{\theta} = \sum_{i \in A_0} \hat{w}_i y_i$$

where

$$\hat{w} = \arg\min_{w} \sum_{i \in A_0} G(w_i) q_i \text{ such that } \sum_{i \in A_0} w_i \mathbf{x}_i = \hat{X}^{(0)}_{GLS} \tag{13}$$

where $\hat{X}^{(0)}_{GLS}$ is the GLS estimate of X for all of the X-variables measured in sample A_0, and G is a generalized entropy function.

4.2 Methodology

In order to have a design consistent estimator that incorporates information from samples $A_0, A_1, \ldots, A_M$, we want to add the debiasing constraints,

$$\sum_{i \in A_0} w_i \mathbf{x}_i q_i = \sum_{i \in U} g(d_i) q_i$$

where $g(x) = \partial G(x)/\partial x$ and d_i are the design weights for sample A_0. This yields the optimization problem:

$$\hat{w} = \arg\min_{w} \sum_{i \in A_0} G(w_i) q_i + \lambda \left(\hat{T} - \sum_{i \in A_0} w_i \mathbf{z}_i \right) \tag{14}$$

where $\hat{T} = (\hat{\mathbf{X}}^{(0)}, \sum_{i \in U} g(d_i) q_i)^T$ and $\mathbf{z}_i = ((\mathbf{x}_i^{(0)})^T, g(d_i) q_i)^T$. Then the estimator is $\hat{Y}_{MS} = \sum_{i \in A_0} w_i y_i$.

A more general approach considers the case in which multiple samples can observe Y. In this case, assume that we have samples $A_1, \ldots A_{M_1}$ for which we observe elements of $\mathbf{X}$ and Y and samples $B_1, \ldots, B_{M_2}$ for which we only observed elements of $\mathbf{X}$. We denote $\mathbf{X}^{(A_m)}$ and $\mathbf{X}^{(B_m)}$ as the columns of $\mathbf{X}$ observed in samples A_m and B_m respectively. Then the final estimator[2] is

$$\hat{Y}_{MSE} = M_1^{-1} \sum_{m=1}^{M_1} \sum_{i \in A_m} \hat{w}_{im} y_i$$

and $\hat{w}_{im}$ solves

$$\hat{w} = \arg\min_{w} \sum_{i \in A_m} G(w_i) q_i + \lambda \left(\hat{T}_m - \sum_{i \in A_m} w_i \mathbf{z}_i \right) \tag{15}$$

[2] This is sort of like stratified sampling with unknown strata.

where $\hat{T} = (\hat{\mathbf{X}}^{(A_m)}, \sum_{i\in U} g(d_{im})q_i)^T$, $\mathbf{z}_i^T = ((\mathbf{x}_i^{(A_m)})^T, g(d_{im})q_i)^T$ and d_{im} are the design weights for A_m.

4.3 Theoretical Results

We prove the theory for the simpler case in which we have samples $A_0, A_1, \dots, A_M$.

Theorem 4 (Design Consistency)**.** *Let $\boldsymbol{\lambda}^*$ be the probability limit of $\hat{\boldsymbol{\lambda}}$. Under regularity conditions,*

$$\hat{Y}_{MS} = \hat{Y}_\ell(\boldsymbol{\lambda}^*, \boldsymbol{\phi}^*) + O_p(N/n_0)$$

where

$$\hat{Y}_\ell(\boldsymbol{\lambda}^*, \boldsymbol{\phi}^*) = \hat{Y}_{DCE} + \left(\hat{T} - \sum_{i\in A_0} \hat{w}_{0i}(\boldsymbol{\lambda}^*)\mathbf{z}_i q_i\right)\boldsymbol{\phi}^*$$

and

$$\boldsymbol{\phi}^* = \left[\sum_{i\in U} \frac{\pi_{0i}\mathbf{z}_i\mathbf{z}_i^T q_i}{g'(d_{0i})}\right]^{-1} \sum_{i\in U} \frac{\pi_{0i}\mathbf{z}_i y_i}{g'(d_{0i})}.$$

Proof. This proof follows from the proof of Theorem 2. The only difference is that we have $\hat{X}_{GLS}$ instead of $\hat{X}_c$, but the final result still holds. □

Since this result holds for each simple case, and our general case is the average of each of the simple cases that we observe, the result holds for Equation (15) as we assume each A_m is independent of each other.

Theorem 5 (Variance Estimation)**.** *Under regularity conditions,*

$$V(\hat{Y}_{MS}) = M_1^{-2}\left(\sum_{m=1}^{M_1}\Big\{(\boldsymbol{\phi}_1^{(m)*})^T \, Var(\hat{\mathbf{X}}_{GLS}^{(m)})(\boldsymbol{\phi}_1^{(m)*})^T + \sum_{i\in U}\sum_{j\in U}\frac{\Delta_{mij}}{\pi_{mi}\pi_{mj}}(y_i - \mathbf{z}_i^{(m)}\boldsymbol{\phi}^{(m)*}q_i)(y_j - \mathbf{z}_j^{(m)}\boldsymbol{\phi}^{(m)*}q_j)\Big\}\right).$$

We can estimate the variance with

$$
\begin{aligned}
\hat{V}(\hat{Y}_{MS}) = M_1^{-2} \Bigg(\sum_{m=1}^{M_1} \Big\{ & (\hat{\boldsymbol{\phi}_1}^{(m)})^T \, \widehat{Var}(\hat{\mathbf{X}}_{GLS}^{(m)}) (\hat{\boldsymbol{\phi}_1}^{(m)})^T \\
& + \sum_{i \in A_m} \sum_{j \in A_m} \frac{\Delta_{mij}}{\pi_{mij}\pi_{mi}\pi_{mj}} (y_i - \mathbf{z}_i^{(m)} \hat{\boldsymbol{\phi}}^{(m)} q_i)(y_j - \mathbf{z}_j^{(m)} \hat{\boldsymbol{\phi}}^{(m)} q_j) \Big\} \Bigg).
\end{aligned}
$$

Proof. This result follows the same argument as Theorem (3). The result holds because each sample A_m is independent from the other $A_{m'}$. Since, the score equation for the regression estimator asserts that $\sum_{i \in A_m} \mathbf{z}_i (y_i - \mathbf{z}_i \hat{\boldsymbol{\phi}}^{(m)}) = \mathbf{0}$, the covariance terms should be approximately zero. □

4.4 Simulation Study 3.1

We first run a simulation for the case in which we only have one sample in which we observe Y. We have the following superpopulation model with $N = 10000$ elements:

$$
\begin{aligned}
X_{1i} &\overset{ind}{\sim} N(2, 1) \\
X_{2i} &\overset{ind}{\sim} Unif(0, 4) \\
X_{3i} &\overset{ind}{\sim} N(0, 1) \\
Z_i &\overset{ind}{\sim} N(0, 1) \\
\varepsilon_i &\overset{ind}{\sim} N(0, 1) \\
Y_i &= 3X_{1i} + 2X_{2i} + \varepsilon_i \\
\pi_{0i} &= \min(\max(\Phi(-x_{1i}), 0.02), 0.9) \\
\pi_{1i} &= n_1/N. \\
\pi_{2i} &= \Phi(x_{2i} - 2)
\end{aligned}
$$

We observe the following columns in each sample

For the sampling mechanism both A_0 and A_1 are selected using a Poisson sample with response probabilities π_{0i} and π_{1i} respectively. The sample A_2 is a simple random sample with $n_1 = 2000$. We compare five different estimators for $\theta = E[Y]$.

Sample	X_1	X_2	X_3	Y
A_0	✓	✓	✓	✓
A_1	✓		✓	
A_2	✓	✓		

1. Horvitz-Thompson estimator (HT): $\hat{Y} = N^{-1} \sum_{i \in A_0} \frac{y_i}{\pi_{0i}}$,

2. Non-nested regression (NNReg): This is the non-nested regression from Equation (9) with only using information from Samples A_0 and A_1,

3. Multi-Source proposed (MSEst): This is the proposed estimator from Equation (14),

4. Multi-Source population (MSPop): This is the proposed estimator with using the true value of T_1 from the population,

5. Multi-Source regression (MSReg): This is the linearized regression estimator from FIXME.

Est	Bias	RMSE	EmpCI	Ttest
HT	-0.0340	0.6825	0.948	1.5766
NNReg	-0.0013	0.0972	0.937	0.4383
MSPop	0.0003	0.0568	0.944	0.1721
MSEst	-0.0037	0.0884	0.943	1.3242
MSReg	-0.0039	0.0887	0.939	1.3792

Table 5: This table shows the results of Simulation Study 2. It displays the Bias, RMSE, empirical 95% confidence interval, and a t-statistic assessing the unbiasedness of each estimator for the estimators: HT, NNReg, MSPop, MSEst, and MSReg.

As expected MSPop has the lowest RMSE because it also uses the population totals. It seems like MSEst and MSReg are almost equivalent, which is good because they are also supposed to be identical. The MSEst estimator outperforms NNReg because it also uses information from A_2, even though it implicitly uses a regression estimator with X_3 as a covariate, which is unnecessary.

4.5 Simulation Study 3.2

Building off of the previous simulation study, we now consider the case for which we observe Y in separate samples. We have the following superpopulation model with $N = 10,000$ elements:

$$\begin{aligned}
X_{1i} &\overset{ind}{\sim} N(2,1) \\
X_{2i} &\overset{ind}{\sim} Unif(0,4) \\
X_{3i} &\overset{ind}{\sim} N(5,1) \\
\varepsilon_i &\overset{ind}{\sim} N(0,1) \\
Y_i &= 3X_{1i} + 2X_{2i} + \varepsilon_i \\
\pi_{A1i} &= \min(\max(\Phi(-x_{1i}), 0.02), 0.9) \\
\pi_{A2i} &= \min(\max(\Phi(x_{3i} - 6), 0.02), 0.95) \\
\pi_{B1i} &= n_1/N \\
\pi_{B2i} &= \Phi(x_{2i} - 2)
\end{aligned}$$

We observe the following columns in each sample

Sample	X_1	X_2	X_3	Y
A_1	✓	✓	✓	✓
A_2	✓		✓	✓
B_1	✓		✓	
B_2	✓	✓		

For the sampling mechanism A_1, A_2 and B_2 are selected using a Poisson sample with response probabilities π_{A1i}, π_{A2i} and π_{B1i} respectively. The B_2 sample is a simple random sample (SRS). We estimate the following methods:

1. Horvitz-Thompson estimator for A_1 (HTA1):

$$\hat{Y}_{HT,A_1} = \sum_{i \in A_1} \frac{y_i}{\pi_{A1i}}.$$

2. Horvitz-Thompson estimator for A_2 (HTA1):

$$\hat{Y}_{HT,A_2} = \sum_{i \in A_2} \frac{y_i}{\pi_{A2i}}.$$

3. Weighted Horvitz-Thompson estimator (WHT):

$$\hat{Y}_{WHT} = \frac{\hat{Y}_{HT,A_1}/\hat{V}_{A_1} + \hat{Y}_{HT,A_2}/\hat{V}_{A_2}}{1/\hat{V}_{A_1} + 1/\hat{V}_{A_2}}$$

   where $\hat{V}_{A_m}$ is the estimated variance of the Horvitz-Thompson estimator $\hat{Y}_{HT,A_m}$.

4. Multi-Source proposed (MSEst): This is the proposed estimator from Equation FIXME.

5. Multi-Source population (MSPop): This is the proposed estimator with using the true value of T_1 from the population,

6. Multi-Source regression (MSReg): This is the linearized regression estimator from FIXME.

Est	Bias	RMSE	EmpCI	Ttest
HTA1	-0.0340	0.6825	0.948	1.5766
HTA2	0.0081	0.4053	0.945	0.6315
WHT	-0.0254	0.3434	0.952	2.3452
MSPop	0.0001	0.0547	0.947	0.0587
MSEst	-0.0040	0.0790	0.941	1.5863
MSReg	-0.0044	0.0802	0.934	1.7258

Table 6: This table shows the results of Simulation Study 2. It displays the Bias, RMSE, empirical 95% confidence interval, and a t-statistic assessing the unbiasedness of each estimator for the estimators: HT, NNReg, MSPop, MSEst, and MSReg.

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.

Hidiroglou, M. (2001). Double sampling. *Survey methodology 27*(2), 143–154.

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.

Kish, L. (1965). *Survey Sampling.* John Wiley & Sons, Inc.

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.

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