Debiased Calibration for Generalized Two-Phase Sampling

Caleb Leedy Jae-Kwang Kim

April 13, 2025

Abstract

Generalized entropy calibration (Gneiting and Raftery, 2007) is a nonparametric method to estimate survey weights without using the design weights in the specified loss function. Kwon et al. (2024) explore generalized entropy calibration to estimate survey weights with known population totals. By using a two-phase sampling framework, we extend this method to allow for estimated population totals. We then generalize the two-phase sampling framework to two additional cases: first, when we have two non-nested samples, and second, when we want to combine three or more samples. We apply our method to the Medicare Current Beneficiary Survey for which we get more efficient estimates by combining this data set with population level controls obtained from the Center of Medicare and Medicaid Services as well as survey data from the U.S. Census' American Community Survey.

1 Introduction

Government surveys in the United States provide the public with diverse information about a range of population characteristics. These include health indicators from the National Health and Nutrition Examination Survey, population and housing data from the American Community Survey, and labor force statistics from the Current Population Survey. All of these surveys are probability surveys. Probability sampling is considered the gold-standard for survey sampling. However, obtaining a large probability sample is getting more difficult because people are more likely to choose not to respond. One solution to this problem is data integration. Instead of analyzing a probability sample by itself, we can incorporate information from other probability samples in order to have an analysis with higher statistical power.

The goal of statistical data integration is to combine information from one survey with the variable of interest to another survey with shared covariates. This integration usually involves one of two contexts: incorporating item level responses or using summary statistics. These mirror the micro and macro approaches described in Yang and Kim (2020) respectively. The main technique to incorporate item level responses is mass imputation. With access to individual item level responses, the traditional approach is to build a predictive model of the variable of interest from the covariates shared between the two surveys and predict the missing outcome variable in the survey without the variable of interest. An early use of mass imputation is found in Breidt et al. (1996) where they use mass imputation to improve estimation in the National Resources Inventory survey. Later, Kim and Rao (2012) use a parametric model to develop a novel resampling technique to construct confidence intervals of population level characteristics. Chen et al. (2022) use a nonparametric model to estimate population characteristics. In the macro approach, one uses summary statistics from other studies to adjust the current estimation. This is commonly done using generalized least squares (GLS). Hidiroglou (2001) apply the GLS technique to non-nested two-phase samples. Merkouris (2004) use GLS with multiple surveys and showed an optimality result. This was later extended to area level domains in Merkouris (2010). The main idea of the

GLS approach is to construct a regression estimator for two-phase samples. In Section 2 of the paper, we explain the basic approach of regression estimation and show how it can be generalized to a calibration estimator.

This paper, however, generalizes the two-phase regression estimator using the debiased calibration technique of Kwon et al. (2024). We incorporate summary statistics of other surveys by using a generalized two-phase sample similar to the GLS method. The first phase sample consists of data from multiple sources, and the second phase sample contains the existing data. Unlike previous data integration methods, however, we also include a pre-chosen function of the second phase design weights, which allows us to develop theory showing that our estimates are asymptotically design unbiased and that we have valid confidence intervals. Moreover, we can show that this technique outperforms traditional regression methods in finite samples.

This paper is organized as follows. Section 2 describes the existing methods to combine probability samples using regression estimators or calibration. Section 3 details the generalized two-phase sampling framework and provides theory for our estimator to be design unbiased and the variance estimation to be valid. In Section 4, we extend this framework first to non-nested two-phase sampling and second to multiple samples. Section 5 provides several simulation studies which show that our method works, and Section 6 applies the method to the Medicare Current Beneficiary Survey.

2 Basic Setup

Consider a finite population, $U = \{1, \ldots, N\}$, of size N with measurement $(\mathbf{x}_i, y_i)$, where $\mathbf{x}_i$ is the realization of the auxiliary variables and y_i is the realization of the study variable of interest for element i in the finite population. The auxiliary variables are cheap to measure but the measurement for the study variable is expensive. Two-phase sampling is a popular method for cost effective sampling. In two-phase sampling, the first phase sample A_1 is selected from the target population to obtain the measurement of $\mathbf{x}_i$ in the sample. The sampling design for the first-phase sample can be denoted by $P(A_1)$. After that, based on the observed auxiliary information in the first-phase sample, the second-phase sample $A_2 \subseteq A_1$ is selected from the first-phase sample to observe y_i. The sampling design for the second-phase sample is denoted as $P(A_2 \mid A_1)$. The goal of two-phase sampling is to construct an estimator of $Y_N = \sum_{i=1}^N y_i$ that uses both the observed information in the Phase 2 sample and also the auxiliary information from $\mathbf{X} \in \mathbb{R}^p$ in the Phase 1 sample. How to combine information efficiently is an important problem in two-phase sampling.

Let $\pi_{1i} = \sum_{A_1 : i \in A_1} \Pr(A_1)$ be the first-order inclusion probability for the first-phase sample and let $\pi_{2i|1} = \sum_{A_2 : i \in A_2} \Pr(A_2 \mid A_1)$ be the first-order inclusion probability for the second-phase sample. We write $d_{1i} = \pi_{1i}^{-1}$ and $d_{2i|1} = \pi_{2i|1}^{-1}$ to denote the design weights. From the first-phase sample, we can compute $\widehat{\mathbf{X}}_1 = \sum_{i \in A_1} d_{1i} \mathbf{x}_i$ as an unbiased estimator of $\mathbf{X}_N = \sum_{i=1}^N \mathbf{x}_i$. Also, from the second-phase sample, we can compute $\left(\widehat{\mathbf{X}}_2, \hat{Y}_2\right) = \sum_{i \in A_2} d_{1i} d_{2i|1} (\mathbf{x}_i, y_i)$ as an unbiased estimator of $(\mathbf{X}_N, Y_N)$. We therefore have three estimators for two parameters. The two-phase regression estimator can be used to combine this information efficiently and it can be expressed as

$$\hat{Y}_{\text{reg,tp}} = \sum_{i \in A_1} d_{1i} \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}_q + \sum_{i \in A_2} d_{1i} d_{2i|1} \left(y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}_q \right) \tag{1}$$

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

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

Regarding the choice of q_i, one can consider a linear regression model

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + e_i \tag{2}$$

where $e_i \sim (0, \sigma^2 c_i)$, where $c_i = c(\mathbf{x}_i)$ is a known function. In this case, $q_i = c_i^{-1}$ is a reasonable choice. One way to understand the two-phase regression estimator in (1) is to use the following decomposition:

$$\hat{Y}_{\text{reg,tp}} = \underbrace{\sum_{i\in A_1} d_{1i}\mathbf{x}_i^\top \hat{\boldsymbol{\beta}}_q}_{\text{Prediction}} + \underbrace{\sum_{i\in A_2} d_{1i} d_{2i|1}(y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}_q)}_{\text{Bias correction}}.$$

The prediction term is obtained by finding the best linear unbiased predictor under the working superpopulation model. The bias correction term makes the two-phase regression estimator asymptotically design unbiased regardless if the underlying the superpopulation model is correct or not. We can find an equivalent estimator in the form $\hat{Y}_{\text{reg,tp}} = \sum_{i\in A_2} d_{1i}\hat{w}_{2i}y_i$. This is useful because it shows that the estimator is linear and it is interesting because the estimated weights $\hat{w}_{2i}$ are the weights closest to the design weights that satisfy the constraints from (3).

$$\sum_{i\in A_2} d_{1i} w_{2i}\mathbf{x}_i = \sum_{i\in A_1} d_{1i}\mathbf{x}_i. \tag{3}$$

Closeness is in an L^2-sense. Overall, the weights $\hat{w}_{2i}$ are the solution to the following optimization problem:

$$\hat{w}_2 = \arg\min_{w_2} \sum_{i\in A_2} (w_{2i} - d_{2i|1})^2 q_i \text{ such that } \sum_{i\in A_2} d_{1i} w_{2i}\mathbf{x}_i = \sum_{i\in A_1} d_{1i}\mathbf{x}_i.$$

This means that $\hat{Y}_{\text{reg,tp}}$ 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 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 $g(w_i, d_i) = \partial G/\partial w$ being continuous with respect to w_i.[1] This generalization includes empirical likelihood estimation, and maximum entropy estimation among others (Schennach, 2007).

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

The Deville and Sarndal (1992) method incorporates the design weights into the loss function, which is the part minimizing the variance. We would rather have bias calibration separate from the minimizing the variance so that we can control each in isolation. Kwon et al. (2024) shows that for a generalized entropy function $G(w)$, including a term of $g(d_{2i|1})$ 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).

However, the method of Kwon et al. (2024) assumes known finite population calibration levels. It 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 this paper, we extend their method to two-phase sampling so that we construct asymptotically valid confidence intervals when using estimated finite population totals derived from the Phase 1 sample.

3 Main Proposal

We follow the approach of Kwon et al. (2024) for the debiased calibration method. Let $G : \mathcal{V} \to \mathbb{R}$ be the generalized entropy function that is strictly convex and differentiable. We consider maximizing

$$H(w_2) = -\sum_{i \in A_2} d_{1i} G(w_{2i}) q_i \tag{4}$$

subject to the following constraints:

$$\sum_{i \in A_2} d_{1i} w_{2i} \mathbf{x}_i = \sum_{i \in A_1} d_{1i} \mathbf{x}_i \tag{5}$$

and

$$\sum_{i \in A_2} d_{1i} w_{2i} g(d_{2i|1}) q_i = \sum_{i \in A_1} d_{1i} g(d_{2i|1}) q_i, \tag{6}$$

where $g(w) = \frac{\partial G}{\partial w}$.

Also known as covariate balancing (Imai and Ratkovic, 2014), constraint (5) ensures that the weighted mean of the covariates, $\mathbf{x}_i$, are the same. This is also the usual calibration constraint motivated from the regression model in (2). Constraint (6) is how we incorporate the design weights. Using $g(w) = \frac{\partial G}{\partial w}$, ultimately ensures that $\hat{w}_{2i}$ converges to d_{2i} in probability. This ensures that design consistency is achieved. The original method of Kwon et al. (2024) only considers known finite population quantities on the right hand side of (5). We extend their method within our two-phase framework to allow for an estimated quantity on the right hand side of (5). To estimate the weights, we solve

$$\hat{w}_2 = \arg\min_{w_2} \sum_{i \in A_2} d_{1i} G(w_{2i}) q_i \text{ such that} \sum_{i \in A_2} d_{1i} w_{2i} \mathbf{z}_i = \sum_{i \in A_1} d_{1i} \mathbf{z}_i \tag{7}$$

where $\mathbf{z}_i = (\mathbf{x}_i, g(d_{2i|1}) q_i)$. The resulting estimator of $\bar{Y}_N$ is

$$\hat{Y}_{\text{DCE}} = N^{-1} \sum_{i \in A_2} d_{1i} \hat{w}_{2i} y_i. \tag{8}$$

3.1 Theoretical Results

We compute the estimated weights, $\hat{w}_{2i}$ from (7) using the method of Lagrange multipliers. We minimize the Lagrangian function

$$L(w_{2i}, \boldsymbol{\lambda}) = \sum_{i \in A_2} d_{1i} G(w_{2i}) q_i + \boldsymbol{\lambda}^\top \left(\sum_{i \in A_1} d_{1i} \mathbf{z}_i - \sum_{i \in A_2} d_{1i} w_{2i} \mathbf{z}_i \right) \tag{9}$$

where $\boldsymbol{\lambda} \in \mathbb{R}^{p+1}$ is a vector of Lagrange multipliers and $\mathbf{z}_i = (\mathbf{x}_i^\top, g(d_{2i|1}) q_i)^\top$. Differentiating with respect to w_{2i} and setting this expression equal to zero, we can obtain

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

where $g(w) = \frac{\partial G}{\partial w}$ and $\hat{\boldsymbol{\lambda}}$ is the solution to

$$\sum_{i \in A_1} d_{1i} \mathbf{z}_i - \sum_{i \in A_2} d_{1i} w_{2i}(\hat{\boldsymbol{\lambda}}) \mathbf{z}_i = 0. \tag{10}$$

Under regularity conditions presented in the appendix, this solution is unique. The following theorem presents a linearization of (7) which leads to a design consistent estimator $\hat{Y}_{\text{DCE}}$ in (8).

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

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

where

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

and

$$\boldsymbol{\phi}^* = \left[\sum_{i \in U} \frac{\pi_{2i|1} \mathbf{z}_i \mathbf{z}_i^\top}{g'(d_{2i|1}) q_i} \right]^{-1} \sum_{i \in U} \frac{\pi_{2i|1} \mathbf{z}_i y_i}{g'(d_{2i|1}) q_i}.$$

See the appendix for the proof of Theorem 1. Notice that this theorem implies that $\hat{Y}_\ell(\boldsymbol{\lambda}^*, \boldsymbol{\phi}^*)$ and $\hat{Y}_{\text{DCE}}$ have the same asymptotic expectation and variance, which means that we can estimate the variance of $\hat{Y}_\ell(\boldsymbol{\lambda}^*, \boldsymbol{\phi}^*)$ to get an estimate of the variance of $\hat{Y}_{\text{DCE}}$. The basic idea of Theorem 1 is that asymptotically the two-phase debiased calibration method is equivalent to a two-phase regression estimator. By Theorem 1, $\hat{Y}_\ell(\boldsymbol{\lambda}^*, \boldsymbol{\phi}^*)$ is design consistent, and the proposed calibration estimator in (7) is asymptotically equivalent to the two-phase regression estimator given by

$$\hat{Y}_{\text{tp,reg}} = \sum_{i \in A_2} d_{1i} d_{2i|1} y_i + \left(\sum_{i \in A_1} d_{1i} \mathbf{z}_i - \sum_{i \in A_2} d_{1i} d_{2i|1} \mathbf{z}_i \right) \hat{\boldsymbol{\phi}},$$

where $\hat{\boldsymbol{\phi}}$ is defined in (11)

$$\hat{\boldsymbol{\phi}} = \left[\sum_{i \in A_2} d_{1i} \frac{\mathbf{z}_i \mathbf{z}_i^\top}{g'(d_{2i|1}) q_i} \right]^{-1} \sum_{i \in A_2} d_{1i} \frac{\mathbf{z}_i y_i}{g'(d_{2i|1}) q_i}. \tag{11}$$

For specific sample designs, we can construct the optimal entropy function $G(\cdot)$. From Theorem 1, the choice of entropy function affects the variance of the debiased calibration estimator through the form of $g'(d_{2i|1})$ in $\boldsymbol{\phi}^*$. Consider a class of generalized regression estimators,

$$\mathbf{B} = \left\{ \sum_{i \in A_1} d_{1i} \mathbf{z}_i^\top \boldsymbol{\gamma} + \sum_{i \in A_2} d_{1i} d_{2i|1} (y_i - \mathbf{z}_i^\top \boldsymbol{\gamma}) : \boldsymbol{\gamma} \in \mathbb{R}^{p+1} \right\}$$

The design optimal estimator is

$$\hat{\theta}_{\ell,\text{opt}} = \sum_{i \in A_1} d_{1i} \mathbf{z}_i^\top \boldsymbol{\phi}_{\text{opt}} + \sum_{i \in A_2} d_{1i} d_{2i|1} (y_i - \mathbf{z}_i^\top \boldsymbol{\phi}_{\text{opt}}) :$$

where

$$\boldsymbol{\phi}_{\text{opt}} = \text{Var}\left(\sum_{i \in A_2} d_{2i|1} \mathbf{z}_i \right)^{-1} \text{Cov}\left(\sum_{i \in A_2} d_{2i|1} \mathbf{z}_i, \sum_{i \in A_2}, d_{2i|1} y_i \right).$$

We assume that both A_1 and A_2 are Poisson samples. Then,

$$\boldsymbol{\phi}_{\text{opt}} = \left\{ \sum_{i \in U} \left(d_{2i|1} - 1 + d_{2i|1}(d_{1i} - 1) \right) \mathbf{z_i}\mathbf{z_i}^\top \right\}^{-1} \sum_{i \in U} \left(d_{2i|1} - 1 + d_{2i|1}(d_{1i} - 1) \right) \mathbf{z_i} y_i.$$

To find the optimal entropy function, we solve the differential equation

$$\frac{1}{d_{2i|1} g'(d_{2i|1})} = d_{2i|1} - 1 + d_{2i|1}(d_{1i} - 1).$$

Using the fact that $(y(xy-1))^{-1} = -y^{-1} + x(xy-1)^{-1}$, the optimal entropy function is

$$G(w_{2i|1}) = -\log(w_{2i|1}) + \log(w_{2i|1} d_{1i} - 1).$$

To construct an asymptotically unbiased estimate of the variance we assume that the sampling designs for A_1 and A_2 are measurable and we define the following joint inclusion probabilities:

$$\Pr(i \in A_1, j \in A_1) = \pi_{1ij}, \text{ and } \Pr(i \in A_2, j \in A_2 \mid i \in A_1, j \in A_1) = \pi_{2ij|1}.$$

The next theorem discusses how to compute an estimate for the variance of $\hat{Y}_{\text{DCE}}$. For the proof see the appendix. The idea behind the proof is to use the two-phase sampling approach to estimate the variance. If the sampling fraction $n_2/n_1 \to 0$ as N increases, then the reverse framework of Fay (1992) allows us to make a good approximation.

Theorem 2 (Variance Estimation)**.** *Define $\hat{\eta}_i = \mathbf{z}_i \hat{\boldsymbol{\phi}} + \delta_i d_{2i|1}(y_i - \mathbf{z}_i \hat{\boldsymbol{\phi}})$. We can construct an unbiased estimate the variance of $\hat{Y}_{\text{DCE}}$ with*

$$\hat{V}_{\text{DCE}} = \sum_{i \in A_1} \sum_{j \in A_1} d_{1ij} \Delta_{1ij} \hat{\eta}_i \hat{\eta}_j + \sum_{i \in A_2} \sum_{j \in A_2} d_{1i} d_{2ij|1} \Delta_{2ij|1} d_{2i|1}(y_i - \mathbf{z}_i \hat{\boldsymbol{\phi}}) d_{2j|1}(y_j - \mathbf{z}_j \hat{\boldsymbol{\phi}})$$

where $\Delta_{1ij} = \pi_{1ij} - \pi_{1i}\pi_{1j}$, $\Delta_{2ij|1} = \pi_{2ij|1} - \pi_{2i|1}\pi_{2j|1}$, $d_{1ij} = \pi_{1ij}^{-1}$, $d_{2ij|1} = \pi_{2ij|1}^{-1}$, $\pi_{1ij} = \Pr(i \in A_1, j \in A_1)$, and $\pi_{1ij} = \Pr(i \in A_2, j \in A_2 | i \in A_1, j \in A_1)$.

4 Extensions

4.1 Non-nested Two-Phase Sampling

We now 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. Like traditional two-phase sampling, we consider the Phase 1 sample, A_1, to consist of observations of $(\mathbf{X}_i)$ and the Phase 2 sample, A_2, to consist of observations of $(\mathbf{X}_i, Y_i)$. We denote the sample size of A_1 as n_1 and the sample size of A_2 as 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 $\mathbf{X}_N$,

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

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

$$\hat{Y}_{\text{NN,reg}} = \hat{Y}_2 + (\hat{\mathbf{X}}_c - \hat{\mathbf{X}}_2)^\top \hat{\boldsymbol{\beta}}_q = \hat{Y}_2 + (\hat{\mathbf{X}}_1 - \hat{\mathbf{X}}_2)^\top W \hat{\boldsymbol{\beta}}_q \tag{12}$$

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

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

From an optimality perspective the choice of using the effective sample size to weight the estimates from A_1 and A_2 is reasonable because the effective sample sizes are often proportional to the variance of the estimates of $\mathbf{X}_N$. While the inverse variance weighted estimate is optimal to combine independent samples for a linear estimate, using the effective sample size approximates this procedure without requiring the actual variance of an estimator to be known. Since the samples A_1 and A_2 are independent,

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

where $\boldsymbol{\beta}_q^*$ is the probability limit of $\hat{\boldsymbol{\beta}}_q$. Like the two-phase sample, the regression estimator in (12) can be viewed as a linear estimator given a particular choice of weights $\hat{w}_{2i}$. These weights are the solution to the calibration equation in (13).

$$\hat{w}_2 = \arg\min_{w_2} \sum_{i \in A_2} (w_{2i} - d_{2i})^2 q_i \text{ such that } \sum_{i \in A_2} w_{2i} \mathbf{x}_i = \hat{\mathbf{X}}_c \tag{13}$$

These estimated weights yield an estimator of $\theta = E[Y]$ of $\hat{Y}_{\text{NN,reg}} = N^{-1} \sum_{i \in A_2} \hat{w}_{2i} y_i$. We can 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.

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 minimize the variance of the estimator while ensuring that the covariate balancing condition is met. This results in $\hat{w}_2$ being the solution to the optimization problem in (14)

$$\hat{w}_2 = \arg\min_{w_2} \sum_{i \in A_2} G(w_{2i}) q_i \text{ where } \sum_{i \in A_2} w_{2i} \mathbf{x}_i = \hat{\mathbf{X}}_c \text{ and } \sum_{i \in A_2} w_{2i} g(d_{2i}) q_i = \sum_{i \in U} g(d_{2i}) q_i \tag{14}$$

for $g(x) = G'(x)$ and a known choice of $q_i \in \mathbb{R}$. The difference between solving (14) and (7) 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 $\mathbf{X}_N$,

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

If $n_{1,\text{eff}}$ and $n_{2,\text{eff}}$ are the effective samples sizes for A_1 and A_2 respectively, then the optimal combined estimate is

$$\hat{\mathbf{X}}_c = (n_{1,\text{eff}} \hat{\mathbf{X}}_1 + n_{2,\text{eff}} \hat{\mathbf{X}}_2)/(n_{1,\text{eff}} + n_{2,\text{eff}})$$

We can construct a non-nested two-phase estimator $\hat{Y}_{\text{NNE}}$ for $\bar{Y}_N$ where $\hat{Y}_{\text{NNE}} = N^{-1} \sum_{i \in A_2} \hat{w}_{2i} y_i$ and $\hat{w}_{2i}$ solves (14). Like the classical two-phase approach, to solve this setup we minimize the Lagrangian,

$$L(w_{2i}, \boldsymbol{\lambda}) = \sum_{i \in A_2} G(w_{2i}) q_i + \boldsymbol{\lambda}^\top \left(\hat{\mathbf{T}} - \sum_{i \in A_2} w_{2i} \mathbf{z}_i \right). \tag{15}$$

where $\boldsymbol{\lambda}$ are the Lagrange multipliers, $\mathbf{z}_i = (\mathbf{x}_i^\top, g(d_{2i})q_i)^\top$, and

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

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

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

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

$$\hat{\mathbf{T}} - \sum_{i \in A_2} w_{2i}(\hat{\boldsymbol{\lambda}})\mathbf{z}_i = 0. \tag{16}$$

Like the previous section we first prove that the estimator is asymptotically equivalent to a regression estimator. This ensures that it is design consistent. The asymptotic equivalence is useful because our estimator now has the same variance as a regression estimator. Using this fact, we provide a result showing how to estimate the variance.

Theorem 3 (Design Consistency)**.** *Let $\boldsymbol{\lambda}^*$ to be the probability limit of $\hat{\boldsymbol{\lambda}}$, under some regularity conditions, $\hat{Y}_{\text{NNE}} = \hat{Y}_{\ell,\text{NNE}}(\boldsymbol{\lambda}^*, \boldsymbol{\phi}^*) + O_p(Nn_2^{-1})$ where*

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

and

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

The proof of this result is very similar to the proof the Theorem 1. The biggest difference is that the total for $\mathbf{X}_N$ is estimated from both samples using $\hat{\mathbf{X}}_c$ instead of $\hat{\mathbf{X}}_{\text{HT}}$ from the Phase 1 sample. By the dual formulation of (15), for any $\boldsymbol{\phi}$

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

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

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

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

$$\hat{Y}_{\mathrm{NNE}}(\hat{\boldsymbol{\lambda}}) = \hat{Y}_{\ell,\mathrm{NNE}}(\boldsymbol{\lambda}^*, \boldsymbol{\phi}^*) + O_p(Nn_2^{-1}).$$

The next theorem shows how to estimate the variance of $\hat{Y}_{\mathrm{NNE}}$. Since we have already proves that the estimator is asymptotically equivalent to a type of regression estimator, we only need to compute the variance of the regression estimator to get the variance of $\hat{Y}_{\mathrm{NNE}}$. The proof of this result is in the appendix.

Theorem 4 (Variance Estimation)**.** *The variance of* $\hat{Y}_{\mathrm{NNE}}$ *is*

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

We can estimate the variance using

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

where $\Delta_{2ij} = \pi_{2ij} - \pi_{2i}\pi_{2j}$, $d_{2ij} = \pi_{2ij}^{-1}$, $\pi_{2ij} = \Pr(i \in A_2, j \in A_2)$, *and*

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

4.2 Multi-Source Two-Phase Sampling

When considering non-nested two-phase sampling, we focused on the case of having two samples. Now, we incorporate more than two independent samples together with a debiasing constraint. We examine the case in which we want to estimate $\theta = E[Y]$ where Y is only observed in one sample.

Consider the setup in which we have independent samples $A_0, A_1, \ldots, A_M$ where Y is observed only in A_0 but $\mathbf{X}$ is observed in all of the samples with variables $\mathbf{X}^{(0)}$ observed in

A_0 and variables $\mathbf{X}^{(m)}$ observed in A_m for each $m = 1, \ldots, M$. Like the non-nested case, we assume that each survey is sampled independently from the same population sampling frame. Some basic examples of data structures that would benefit from the multi-source two-phase sampling can be found in Table 1.

	Example 1			Example 2		
Sample	X_1	X_2	Y	X_1	X_2	Y
A_0	✓	✓	✓	✓		✓
A_1	✓			✓	✓	
A_2		✓			✓	

Table 1: This tables shows two examples configuations of multi-source sampling. In each example, A_0 contains the variable of interest Y. For Example 1, the covariates in A_1 and A_2 are both contained in A_0. In Example 2, sample A_2 improves the estimation of Y via improving the estimation of X_1 in A_1.

The traditional multi-source approach (Kim, 2024) is to use GLS to obtain an optimal estimator of $\mathbf{X}_N$ from the samples $A_1, \ldots, A_M$. 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

$$\hat{w} = \arg\min_{w} \sum_{i \in A_0} G(w_i) q_i \text{ such that } \sum_{i \in A_0} w_i \mathbf{x}_i = \hat{\mathbf{X}}_{\text{GLS}}^{(0)} \tag{17}$$

where $\hat{\mathbf{X}}_{\text{GLS}}^{(0)}$ is the GLS estimate of $\mathbf{X}$ for all of the X-variables measured in sample A_0, and $G(\cdot)$ is a generalized entropy function. In this setup, all of the data integration occurs when estimating $\hat{\mathbf{X}}_{\text{GLS}}$, which is used as a calibration constraint. 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 constraint from (18) because it ensures design consistency of the estimated weights

$$\sum_{i \in A_0} w_i g(d_i) q_i = \sum_{i \in U} g(d_i) q_i \tag{18}$$

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

$$\hat{w}_0 = \arg\min_{w_0} \sum_{i \in A_0} G(w_{0i}) q_i \text{ such that } \widehat{\mathbf{T}} = \sum_{i \in A_0} w_{0i} \mathbf{z}_i. \tag{19}$$

where $\widehat{\mathbf{T}} = ((\hat{\mathbf{X}}^{(0)}_{\mathrm{GLS}})^\top, \sum_{i \in U} g(d_i) q_i)^\top$, and $\mathbf{z}_i = ((\mathbf{x}_i^{(0)})^\top, g(d_i) q_i)^\top$. Then the estimator is $\hat{Y}_{\mathrm{MS}} = \sum_{i \in A_0} \hat{w}_{0i} y_i$ and we can construct this estimate minimizing the Lagrangian:

$$L(w_0, \boldsymbol{\lambda}) = \sum_{i \in A_0} G(w_{0i}) q_i + \boldsymbol{\lambda}^\top \left(\widehat{\mathbf{T}} - \sum_{i \in A_0} w_{0i} \mathbf{z}_i \right)$$

where $\boldsymbol{\lambda}$ are the Lagrange multipliers from (19). We now show that $\hat{Y}_{\mathrm{MS}}$ is asymptotically equivalent to a regression estimator. This linearization technique shows design consistency.

Theorem 5 (Design Consistency)**.** *Let $\boldsymbol{\lambda}^*$ be the probability limit of $\hat{\boldsymbol{\lambda}}$. Under regularity conditions,*

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

where

$$\hat{Y}_{\ell,\mathrm{MS}}(\boldsymbol{\lambda}^*, \boldsymbol{\phi}^*) = \sum_{i \in A_0} d_{0i} y_i + \left(\widehat{\mathbf{T}} - \sum_{i \in A_0} d_{0i} \mathbf{z}_i \right)^\top \boldsymbol{\phi}^*$$

and

$$\boldsymbol{\phi}^* = \left[\sum_{i \in U} \frac{\pi_{0i} \mathbf{z}_i \mathbf{z}_i^\top q_i}{g'(d_{0i})} \right]^{-1} \sum_{i \in U} \frac{\pi_{0i} \mathbf{z}_i y_i}{g'(d_{0i})}.$$

This proof follows from the proof of Theorem 3. The only difference is that we have $\hat{\mathbf{X}}_{\mathrm{GLS}}$ instead of $\hat{\mathbf{X}}_c$. Since we have shown equivalence between our estimator and a regression estimator we can estimate the variance by giving the variance of the regression estimator. The next theorem gives a method for how to estimate the variance of $\hat{Y}_{\mathrm{MS}}$.

Theorem 6 (Variance Estimation)**.** *Under regularity conditions,*

$$V(\hat{Y}_{\mathrm{MS}}) = (\boldsymbol{\phi}^*)^\top Var(\hat{\mathbf{X}}^{(0)}_{\mathrm{GLS}})(\boldsymbol{\phi}^*) + \sum_{i \in U} \sum_{j \in U} \Delta_{0ij} d_{0i} (y_i - \mathbf{z}_i \boldsymbol{\phi}^* q_i) d_{0j} (y_j - \mathbf{z}_j \boldsymbol{\phi}^* q_j).$$

We can estimate the variance with

$$\hat{V}(\hat{Y}_{\mathrm{MS}}) = (\hat{\boldsymbol{\phi}})^\top \hat{Var}(\hat{\mathbf{X}}^{(0)}_{\mathrm{GLS}})(\hat{\boldsymbol{\phi}}) + \sum_{i \in A_0} \sum_{j \in A_0} d_{0ij} \Delta_{0ij} d_{0i} (y_i - \mathbf{z}_i \hat{\boldsymbol{\phi}} q_i) d_{0j} (y_j - \mathbf{z}_j \hat{\boldsymbol{\phi}} q_j)$$

where $\Delta_{0ij} = \pi_{0ij} - \pi_{0i}\pi_{0j}$, $d_{0ij} = \pi_{0ij}^{-1}$, and $\pi_{0ij} = \Pr(i \in A_0, j \in A_0)$.

The result follows the same argument as Theorem 4. The result holds because each sample A_m is independent from the other $A_{m'}$.

5 Simulation Studies

5.1 Simulation Study 1

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

$$
\begin{aligned}
X_{1i} &\stackrel{ind}{\sim} N(2,1) \\
X_{2i} &\stackrel{ind}{\sim} \text{Unif}(0,4) \\
X_{3i} &\stackrel{ind}{\sim} N(0,1) \\
\varepsilon_i &\stackrel{ind}{\sim} N(0,1) \\
Y_i &= 3X_{1i} + 2X_{2i} + \delta 0.5 X_{3i} + \varepsilon_i \\
\pi_{1i} &= n_1/N \\
\pi_{2i|1} &= \max(\min(\Phi_3(x_{3i} - 1), 0.7), 0.02).
\end{aligned}
$$

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 occurring under Poisson sampling. The Phase 1 sample has $n_1 = 1200$ and the Phase 2 sample has an expected size of $E[n_2] \approx 2588$. In the Phase 1 sample, we observe (X_1, X_2) while in the Phase 2 sample we observe (X_1, X_2, Y). The parameter $\delta \in \{0, 1\}$ controls the effect of model misspecification in the simulation. Let $\mathbf{x}_i = (1, x_{1i}, x_{2i})^\top$. We compare the proposed method for the parameter $\bar{Y}_N$ with four approaches:

1. π^*-estimator (PiStar): $\hat{Y}_{\pi^*} = N^{-1} \sum_{i \in A_2} \frac{y_i}{\pi_{1i}\pi_{2i|1}}$,
2. Two-Phase Regression estimator (Reg): $\hat{Y}_{\text{reg}} = \sum_{i \in A_1} \frac{\mathbf{x}_i' \widehat{\boldsymbol{\beta}}}{\pi_{1i}} + \sum_{i \in A_2} \frac{1}{\pi_{1i}\pi_{2i|1}} (y_i - \mathbf{x}_i' \widehat{\boldsymbol{\beta}})$ where $\widehat{\boldsymbol{\beta}} = \left(\sum_{i \in A_2} \mathbf{x}_i \mathbf{x}_i'\right)^{-1} \sum_{i \in A_2} \mathbf{x}_i y_i$.
3. Debiased Calibration with Population Constraints (EstPop): This solves

$$\arg\min_{w_{2|1}} \sum_{i \in A_2} d_{1i} G(w_{2i}) \text{ such that } \sum_{i \in A_2} d_{1i} w_{2i} \mathbf{z}_i = \sum_{i \in U} \mathbf{z}_i. \tag{20}$$

4. Debiased Calibration with Estimated Population Constraints (Est): This solves (7) with $q_i = 1$.

We run the simulation $B = 1000$ 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. We also include the Monte Carlo variance of the estimated value, ($V_{MC} = (B-1)^{-1} \sum_{b=1}^{B} (\hat{Y}^{(b)} - \bar{Y}_N)^2$), the mean of the estimated variance, ($\bar{\hat{V}} = B^{-1} \sum_{b=1}^{B} \hat{V}^{(b)}$), and the relative bias of the estimated variance, ($(V_{MC} - \bar{\hat{V}})/V_{MC}$). The results are in Table 2 and Table 3.

Est	Bias	SE	RMSE	EmpCI	Ttest	MCVar	EstVar	RelBias
Est	0.003	0.135	0.134	0.945	0.70	0.018	0.017	-0.074
EstPop	0.000	0.082	0.082	0.975	0.07	0.007	0.009	0.317
PiStar	-0.031	0.829	0.829	0.943	1.17	0.688	0.692	0.006
Reg	0.004	0.135	0.135	0.950	0.86	0.018	0.018	-0.031

Table 2: This table shows the results of Simulation Study 1 with $\delta = 0$. It displays the Bias, RMSE, empirical 95% confidence interval, a t-statistic assessing the unbiasedness, the Monte Carlo variance, mean estimated variance and relative bias of the variance estimator for the estimators: PiStar, Reg, EstPop, and Est.

Est	Bias	SE	RMSE	EmpCI	Ttest	MCVar	EstVar	RelBias
Est	0.007	0.136	0.136	0.941	1.56	0.019	0.017	-0.072
EstPop	0.004	0.085	0.085	0.973	1.48	0.007	0.009	0.320
PiStar	-0.028	0.794	0.794	0.941	1.13	0.631	0.635	0.007
Reg	0.011	0.153	0.154	0.940	2.19	0.024	0.024	-0.001

Table 3: This table shows the results of Simulation Study 1 with $\delta = 1$. It displays the Bias, RMSE, empirical 95% confidence interval, a t-statistic assessing the unbiasedness, the Monte Carlo variance, mean estimated variance and relative bias of the variance estimator for the estimators: PiStar, Reg, EstPop, and Est.

5.2 Simulation Study 2

We run a simulation testing the non-nested extension. This is very similar to Simulation 1. We have the following simulation setup:

$$
\begin{aligned}
X_{1i} &\overset{ind}{\sim} N(2,1) \\
X_{2i} &\overset{ind}{\sim} \mathrm{Unif}(0,4) \\
X_{3i} &\overset{ind}{\sim} N(0,1) \\
\varepsilon_i &\overset{ind}{\sim} N(0,1) \\
Y_i &= 3X_{1i} + 2X_{2i} + \delta X_{3i}\varepsilon_i \\
\pi_{1i} &= n_1/N \\
\pi_{2i} &= \max(\min(\Phi_3(X_{3i} - 2.5), 0.9), 0.01).
\end{aligned}
$$

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 about 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). If $\delta = 0$, then there is no model misspecification. However, if $\delta = 1$, then there is some model misspecification. We estimate the parameter $\bar{Y}_N$ with four approaches:

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

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

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

3. Debiased Calibration with Population Constraints (EstPop): This solves

$$\hat{w}_2 = \arg\min_{w_2} \sum_{i \in A_2} G(w_{2i})$$
$$\text{such that } \sum_{i \in A_2} w_{2i}\mathbf{x}_i = \sum_{i \in U} \mathbf{x}_i \text{ and } \sum_{i \in A_2} w_{2i} g(d_{2i}) = \sum_{i \in U} g(d_{2i})$$

4. Debiased Calibration with Estimated Population Constraints (Est): This solves (14) 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 $B = 1000$ 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. We also include the Monte Carlo variance of the estimated value, ($V_{MC} = (B-1)^{-1} \sum_{b=1}^{B} (\hat{Y}^{(b)} - \bar{Y}_N)^2$), the mean of the estimated variance, ($\bar{\hat{V}} = B^{-1} \sum_{b=1}^{B} \hat{V}^{(b)}$), and the relative bias of the estimated variance, ($(V_{MC} - \bar{\hat{V}})/V_{MC}$). The results are in Table 4 and Table 5.

Est	Bias	SE	RMSE	EmpCI	Ttest	MCVar	EstVar	RelBias
Est	0.000	0.123	0.123	0.954	0.01	0.015	0.016	0.040
EstPop	0.000	0.053	0.053	0.943	0.30	0.003	0.003	0.023
HT	0.024	0.565	0.566	0.954	1.33	0.320	0.325	0.017
Reg	0.000	0.123	0.123	0.956	0.11	0.015	0.016	0.057

Table 4: This table shows the results of Simulation Study 2 with $\delta = 0$. It displays the Bias, RMSE, empirical 95% confidence interval, a t-statistic assessing the unbiasedness, the Monte Carlo variance, mean estimated variance and relative bias of the variance estimator for the estimators: HT, Reg, EstPop, and Est.

Est	Bias	SE	RMSE	EmpCI	Ttest	MCVar	EstVar	RelBias
Est	-0.001	0.124	0.124	0.961	0.13	0.015	0.017	0.132
EstPop	0.000	0.066	0.066	0.937	0.14	0.004	0.004	-0.001
HT	0.019	0.529	0.529	0.956	1.14	0.280	0.285	0.017
Reg	-0.003	0.127	0.127	0.965	0.67	0.016	0.020	0.212

Table 5: This table shows the results of Simulation Study 2 with $\delta = 1$. It displays the Bias, RMSE, empirical 95% confidence interval, a t-statistic assessing the unbiasedness, the Monte Carlo variance, mean estimated variance and relative bias of the variance estimator for the estimators: HT, Reg, EstPop, and Est.

5.3 Simulation Study 3

This simulation study tests the multi-source extension. We have the following superpopulation model with $N = 10,000$ elements:

$$
\begin{aligned}
X_{1i} &\overset{ind}{\sim} N(2,1) \\
X_{2i} &\overset{ind}{\sim} \text{Unif}(0,4) \\
X_{3i} &\overset{ind}{\sim} N(5,1) \\
Z_i &\overset{ind}{\sim} N(0,1) \\
\varepsilon_i &\overset{ind}{\sim} N(0,1) \\
Y_i &= 3X_{1i} + 2X_{2i} + \delta Z_i + \varepsilon_i \\
\pi_{0i} &= \min(\max(\Phi(Z_i - 2), 0.02), 0.9) \\
\pi_{1i} &= n_1/N \\
\pi_{2i} &= \Phi(X_{2i} - 2)
\end{aligned}
$$

Like the previous simulation studies, when $\delta = 1$, there is model misspecification for the outcome model because we observe the following columns in each sample:

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

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

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 (14) with only using information from Samples A_0 and A_1,
3. Multi-Source proposed (Est): This is the proposed estimator from (19), and
4. Multi-Source population (EstPop): This is the proposed estimator with using the true value of T_1 from the population.

The simulation results are displayed in Table 6 and Table 7.

Est	Bias	SE	RMSE	EmpCI	Ttest	MCVar	EstVar	RelBias
Est	-0.004	0.091	0.091	0.943	1.55	0.008	0.008	-0.087
EstPop	-0.003	0.058	0.058	0.932	1.40	0.003	0.003	-0.073
HT	-0.006	0.596	0.596	0.951	0.34	0.356	0.353	-0.008
NNReg	-0.005	0.099	0.099	0.941	1.45	0.010	0.009	-0.080

Table 6: This table shows the results of Simulation Study 3 with $\delta = 0$. It displays the Bias, RMSE, empirical 95% confidence interval, a t-statistic assessing the unbiasedness, the Monte Carlo variance, mean estimated variance and relative bias of the variance estimator for the estimators: HT, NNReg, EstPop, and Est.

Est	Bias	SE	RMSE	EmpCI	Ttest	MCVar	EstVar	RelBias
Est	-0.002	0.097	0.097	0.938	0.63	0.009	0.009	-0.051
EstPop	0.000	0.068	0.068	0.945	0.07	0.005	0.004	-0.049
HT	-0.006	0.569	0.569	0.947	0.34	0.324	0.321	-0.008
NNReg	-0.002	0.107	0.106	0.974	0.49	0.011	0.015	0.345

Table 7: This table shows the results of Simulation Study 3 with $\delta = 1$. It displays the Bias, RMSE, empirical 95% confidence interval, a t-statistic assessing the unbiasedness, the Monte Carlo variance, mean estimated variance and relative bias of the variance estimator for the estimators: HT, NNReg, EstPop, and Est.

As expected, EstPop has the lowest RMSE because it also uses the population totals. The Est 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.

6 Real Data Analysis

We apply our method to Medicare Current Beneficiary Survey (MCBS). The MCBS is an ongoing representative national survey of the population of people on Medicare in the United States, and in 2020, NORC started issuing an MCBS supplement to identify the impact of COVID-19 on the Medicare population (NORC, 2024). This analysis looks at whether respondents had obtained at least one dose of a COVID-19 vaccine.

Assessing the vaccine rate of Medicare patients is important from a public health perspective. Older people[2] are known to have a higher all-cause mortality from COVID-19 (Bonanad et al., 2020). Thus, understanding what causes older people to get vaccinated can help save lives and reduce medical costs.

The MCBS is a rotating panel survey in which participants are interviewed up to three times within a four-year period (CMS, 2021). The participants are selected from Medicare enrollment data and enter a cohort in the fall round of the survey. They then respond to the winter, summer, and fall rounds of the survey for three consecutive years before exiting the survey in the winter round of year four (CMS, 2021). Each year a new cohort is added. Participants are selected based on a three-stage cluster sample design. The first stage of the sample consists of major metropolitan areas and groups of rural counties (CMS, 2021). The second stage are census tracts within the primary sampling units. Finally, Medicare beneficiaries are selected using a stratified systematic sample with random starts (CMS, 2021). Medicare beneficiaries are selected from strata based on age and whether the beneficiary is Hispanic. Strata with Hispanic beneficiaries and strata with beneficiaries over the age of 85 and under the age of 65 are oversampled. The COVID supplement was added to the traditional MCBS survey in March of 2020 to assess the impact of COVID

[2]Medicare is available to people ages 65 years and older and select designated populations between the ages of 18 and 65.

on the Medicare population (CMS, 2022). These questions about vaccines were added to participants who were enrolled in the Summer 2021 cohort.

We combine the 2021 Summer MCBS survey with data from the 2021 American Community Survey (ACS) issued by the U.S. Census Bureau along with administrative data about people on Medicare from the Center of Medicare and Medicaid Services (CMS). Like the MCBS survey, the ACS data contains information about the overall distribution of sex, race and ethnicity, education, and marital status for people on Medicare. We combine this information with official totals of people on Medicare by sex from the CMS. A complete display of the data available can be seen in Table 6.

Data	Sex	Race and Ethnicity	Education	Marital Status	COVID Vaccine Dose
MCBS	✓	✓	✓	✓	✓
ACS	✓	✓	✓	✓	
CMS	✓				

Table 8: This table shows the data available to us in each data set. A check mark in a particular column indicates that the data set listed in the left column contains information about the participant characteristic in the given column.

To make the results between the ACS and MCBS compatible, we recode the ACS variables with the map found in Table 6.

Variable	Consolidated ACS Variables
Education	SCHL: 01-15, 16-19, 20-24
Marital Status	MAR: 1, 2, 3-4, 5

Table 9: This table shows how we consolidated ACS variables. The right column displays the ACS category code and the values used in a particular category to match a MCBS category. For example, SCHL: 01-15 means that the values of 01 - 15 of the SCHL ACS column were combined to match an individual MCBS value.

While the MCBS data is largely free of missing values, there are a small number that need to be imputed for the Y variable about whether someone received at least one COVID vaccination dose. This MCBS data was updated to include information about whether a participant had received a second vaccine dose. Later, the survey included the total vaccine doses patients had received. If the participant had received at least one more dose and their answer was missing (because they refused to answer, did not know, or was missing) the missing value was imputed to say that they yes, they had received a COVID vaccine dose. Otherwise, the missing value was imputed to say no, the participant had not received a COVID vaccine dose.

We analyze the data with three approaches: a Horvitz-Thompson method, a regression algorithm, and our preposed multi-source two-phase sampling technique. The Horvitz-Thompson estimator uses the design weights, the MCBS observation whether an individual received a COVID-19 vaccine dose, and the population total from the CMS data about the total number of people on Medicare in 2021. The regression estimator uses all of this and a population level total for the number of males. Our multi-source algorithm also includes information from the ACS about education and marital status. Due to the fact that we do not observe the probability that an individual will be selected into the MCBS survey within the population of all Medicare patients, we have to estimate the term $\sum_{i \in U} g(d_i)$ from (18). This technique is adopted from Kwon et al. (2024) and described in more details in the appendix. The results of the real data analysis are found in Table 6.

Method	Point Estimate	Standard Error
HT	0.817	0.01223
Reg	0.877	0.00439
Est	0.863	0.00436

Table 10: This table shows the point estimate and standard error of the percentage of people on Medicare who had at least one COVID-19 vaccine in 2021 as estimated by a Horvitz-Thompson estimator (HT), a regression estimator (Reg) and the proposed multi-source estimator (Est).

7 Conclusion

Overall, using debiased calibration for generalized two-phase sampling seems to be a promising approach for combining multiple samples efficiently. Not only can this method be used to combine surveys focusing on the same population frame, but because it does not require individual responses levels from the supplemental surveys, this method can also be used to construct domain level estimates as exemplified in the real data analysis of Medicare patients.

References

Bonanad, C., Garcia-Blas, S., Tarazona-Santabalbina, F., Sanchis, J., Bertomeu-Gonzalez, V., Facila, L., Ariza, A., Nunez, J., and Cordero, A. (2020). The effect of age on mortality in patients with covid-19: a meta-analysis with 611,583 subjects. *Journal of the American Medical Directors Association*, 21(7):915–918.

Breidt, F. J., McVey, A., and Fuller, W. A. (1996). Two-phase estimation by imputation. *Journal of the Indian Society of Agricultural Statistics*, 49:79–90.

Chen, S., Yang, S., and Kim, J. K. (2022). Nonparametric mass imputation for data integration. *Journal of survey statistics and methodology*, 10(1):1–24.

CMS (2021). Mcbs tutorial.

CMS (2022). Mcbs advanced tutorial on the covid-19 supplement data.

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

Fay, R. E. (1992). When are inferences from multiple imputation valid?.

Gneiting, T. and Raftery, A. E. (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.

Imai, K. and Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):243–263.

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

Kim, J. K. and Rao, J. N. (2012). Combining data from two independent surveys: a model-assisted approach. *Biometrika*, 99(1):85–100.

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

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

Merkouris, T. (2004). Combining independent regression estimators from multiple surveys. *Journal of the American Statistical Association*, 99(468):1131–1139.

Merkouris, T. (2010). Combining information from multiple surveys by using regression for efficient small domain estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(1):27–48.

NORC (2024). Medicare current beneficiary survey (mcbs) covid-19 supplements.

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

Schennach, S. M. (2007). Point estimation with exponentially tilted empirical likelihood. *The Annals of Statistics*, 35:634–672.

Shao, J. and Steel, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, 94(445):254–265.

Yang, S. and Kim, J. K. (2020). Statistical data integration in survey sampling: A review. *Japanese Journal of Statistics and Data Science*, 3:625–650.

A Appendix: Regularity Conditions

We need the following regularity conditions to prove Theorem 1.

A1 : $\mathcal{V} \in \mathbb{R}$ is open,

A2 : $G(w)$ is strictly convex and twice continuously differentiable on $\mathcal{V}$ with a bounded second derivative,

A3 : There exist positives constants $c_1, c_2 \in \mathcal{V}$ such that $c_1 < \pi_{2i|1} < c_2$ for all π_i,

A4 : For $\Delta_{2ij|1} = \pi_{2ij|1} - \pi_{2i|1}\pi_{2j|1}$,

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

A5 : The matrix $\mathbf{\Sigma}_{\mathbf{z}} := \lim_{N\to\infty} N^{-1} \sum_{i\in U} \mathbf{z}_i \mathbf{z}_i^\top$ exists and is positive definite,

A6 : The average fourth moment of $(\mathbf{x}_i^\top, y_i)^\top$ is finite such that

$$\limsup_{N\to\infty} N^{-1} \sum_{i=1}^{N} (\mathbf{x}_i^\top, y_i)^\top < \infty,$$

A7 : $||\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^*|| = O_p(n_2^{-1/2})$

Assumption A1 is needed for the differentiability of $G(w)$ in A2. Assumption A2 guarentees a unique solution for $\hat{\boldsymbol{\lambda}}$ and ensures that $g(w)$ is differentiable. The boundedness part of the assumption ensures that $\frac{\partial}{\partial \boldsymbol{\lambda}} g^{-1}(\boldsymbol{\lambda}^\top \mathbf{z})$ is also bounded. Conditions A3 and A4 are standard conditions which control the first-order and second-order inclusion probability for the second phase sample. These conditions are met by many traditional survey designs. Assumptions A5 and A6 are technical conditions and common in the survey literature. Assumption A7 is added to simplify the proof. We could remove this condition and follow the approach of Theorem 1 in Kwon et al. (2024).

A.1 Proof of Theorem 1

In this proof, we derive the solution to (7) and show that it is asymptotically equivalent to a regression estimator. Using the method of Lagrange multipliers, to solve (7) we need to minimize the Lagrangian in (9). The first order conditions show that at the solution,

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

Hence, $\hat{w}_{2i}(\hat{\boldsymbol{\lambda}}) = g^{-1}(\hat{\boldsymbol{\lambda}}^\top \mathbf{z}_i)$ where $\hat{\boldsymbol{\lambda}}$ is determined by (10). When the sample size gets large, we have $\hat{w}_{2i}(\hat{\boldsymbol{\lambda}}) \to d_{2i|1}$ in probability because both terms in (10) must become equal by the Central Limit Theorem. Hence, $\hat{\boldsymbol{\lambda}} \to \boldsymbol{\lambda}^*$ in probability and $\boldsymbol{\lambda}^* = (\mathbf{0}^\top, 1)^\top$. Using the linearization technique of Randles (1982), we can construct a regression estimator,

$$\hat{Y}_\ell(\hat{\boldsymbol{\lambda}}, \boldsymbol{\phi}) = \sum_{i \in A_2} d_{1i} \hat{w}_{2i} y_i + \left(\sum_{i \in A_1} d_{1i} \boldsymbol{z}_i - \sum_{i \in A_2} d_{1i} \hat{w}_{2i} \mathbf{z}_i \right)^\top \boldsymbol{\phi}.$$

Notice that $\hat{Y}_\ell(\hat{\boldsymbol{\lambda}}, \boldsymbol{\phi}) = \sum_{i \in A_2} d_{1i} \hat{w}_{2i} y_i$ for all $\boldsymbol{\phi}$. We choose $\boldsymbol{\phi}^*$ such that

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

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

$$\begin{aligned}
\boldsymbol{\phi}^* &= E\left[\sum_{i \in A_2} \frac{d_{1i} \mathbf{z}_i \mathbf{z}_i^\top}{g'(d_{2i|1}) q_i} \right]^{-1} E\left[\sum_{i \in A_2} \frac{d_{1i} \mathbf{z}_i y_i}{g'(d_{2i|1}) q_i} \right] \\
&= \left[\sum_{i \in U} \frac{\pi_{2i|1} \mathbf{z}_i \mathbf{z}_i^\top}{g'(d_{2i|1}) q_i} \right]^{-1} \left[\sum_{i \in U} \frac{\pi_{2i|1} \mathbf{z}_i y_i}{g'(d_{2i|1}) q_i} \right]
\end{aligned}$$

Thus, the linearization estimator is

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

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

A.2 Proof of Theorem 2

Proof. By Theorem 1, $\hat{Y}_{\text{DCE}}$ is asymptotically equivalent to $\hat{Y}_\ell$. Notice that we can define

$$\eta_i^* = \mathbf{z}_i\boldsymbol{\phi}^* + \delta_i d_{2i|1}(y_i - \mathbf{z}_i^\top \boldsymbol{\phi}^*)$$

where δ_i is one if $i \in A_2$ and zero otherwise. Then,

$$\hat{Y}_\ell = \sum_{i \in A_1} d_{1i}\eta_i^*.$$

Using the reverse framework of Fay (1992) and Shao and Steel (1999), and noting that $\Delta_{1ij} = \pi_{1ij} - \pi_{1i}\pi_{1j}$ and $\Delta_{2ij|1} = \pi_{2ij|1} - \pi_{2i|1}\pi_{2j|1}$,

$$\begin{aligned}
\text{Var}(\hat{Y}_{\ell,\text{DCE}}) &= \text{Var}\left(\sum_{i \in A_1} d_{1i}\eta_i^*\right) \\
&= \text{Var}\left(E\left[\sum_{i \in A_1} d_{1i}\eta_i^* \mid A_2\right]\right) + E\left[\text{Var}\left(\sum_{i \in A_1} d_{1i}\eta_i^* \mid A_2\right)\right] \\
&= \text{Var}\left(\sum_{i \in U} \eta_i^*\right) + E\left[\sum_{i \in U}\sum_{j \in U} \Delta_{1ij} d_{1i}\eta_i^* w_{1j}\eta_j^*\right]
\end{aligned}$$

This means that we can estimate the variance of $\hat{Y}_{\text{DCE}}$ with

$$\hat{V}_{\text{DCE}} = \sum_{i \in A_1}\sum_{j \in A_1} d_{1ij}\Delta_{1ij}\hat{\eta}_i\hat{\eta}_j + \sum_{i \in A_2}\sum_{j \in A_2} d_{1i}d_{2ij|1}\Delta_{2ij|1}d_{2i|1}(y_i - \mathbf{z}_i\hat{\boldsymbol{\phi}})d_{2j|1}(y_j - \mathbf{z}_j\hat{\boldsymbol{\phi}}).$$

□

A.3 Proof of Theorem 4

Proof. From Theorem 3, we know that $\hat{Y}_{\text{NNE}}(\hat{\lambda}) = \hat{Y}_{\ell,\text{NNE}}(\boldsymbol{\lambda}^*, \boldsymbol{\phi}^*) + O_p(Nn_2^{-1})$. Hence, the variance of $\hat{Y}_{\text{NNE}}(\hat{\boldsymbol{\lambda}})$ is

$$
\begin{aligned}
\text{Var}(\hat{Y}_{\text{NNE}}(\hat{\boldsymbol{\lambda}})) &\doteq \text{Var}(\hat{Y}_{\ell,\text{NNE}}(\boldsymbol{\lambda}^*, \boldsymbol{\phi}^*)) \\
&= \text{Var}\left(\sum_{i\in A_2} \hat{w}_{2i}(\boldsymbol{\lambda}^*)y_i + \left(\mathbf{T} - \sum_{i\in A_2} \hat{w}_{2i}(\boldsymbol{\lambda}^*)\mathbf{z}_i\right)^\top \boldsymbol{\phi}^*\right) \\
&= (\boldsymbol{\phi}_1^*)^\top \text{Var}(\hat{\mathbf{X}}_c)\boldsymbol{\phi}_1^* + \sum_{i\in U}\sum_{j\in U} \Delta_{2ij} d_{2i}(y_i - \mathbf{z}_i^\top \boldsymbol{\phi}^*) d_{2j}(y_j - \mathbf{z}_j^\top \boldsymbol{\phi}^*) \\
&\quad + 2\text{Cov}\left(\hat{\mathbf{X}}_c^\top \boldsymbol{\phi}_1^*, \sum_{i\in A_2} d_{2i}(y_i - \mathbf{z}_i^\top \boldsymbol{\phi}^*)\right) \\
&= (\boldsymbol{\phi}_1^*)^\top \text{Var}(\hat{\mathbf{X}}_c)\boldsymbol{\phi}_1^* + \sum_{i\in U}\sum_{j\in U} \Delta_{2ij} d_{2i}(y_i - \mathbf{z}_i^\top \boldsymbol{\phi}^*) d_{2j}(y_j - \mathbf{z}_j^\top \boldsymbol{\phi}^*) \\
&\quad + (1-W)\left(\sum_{i\in U}\sum_{j\in U} \Delta_{2ij} d_{2i}\mathbf{x}_i d_{2j}(y_j - \mathbf{z}_j^\top \boldsymbol{\phi}_1^*)\right)^\top \boldsymbol{\phi}_1^*
\end{aligned}
$$

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:

$$
\begin{aligned}
\hat{V}_{\text{NNE}} = (\hat{\boldsymbol{\phi}}_1)^\top \text{Var}(\hat{\mathbf{X}}_c)\hat{\boldsymbol{\phi}}_1 &+ \sum_{i\in A_2}\sum_{j\in A_2} d_{2ij}\Delta_{2ij} d_{2i}(y_i - \mathbf{z}_i^\top \hat{\boldsymbol{\phi}}) d_{2j}(y_j - \mathbf{z}_j^\top \hat{\boldsymbol{\phi}}) \\
&+ (1-W)\left(\sum_{i\in A_2}\sum_{j\in A_2} d_{2ij}\Delta_{2ij} d_{2i}\mathbf{x}_i d_{2j}(y_j - \mathbf{z}_j^\top \hat{\boldsymbol{\phi}}_1)\right)^\top \hat{\boldsymbol{\phi}}_1
\end{aligned}
$$

where

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

□

B Estimating the Population Weights

In (7), (14), and (18), we see the need for knowing $\sum_{i\in U} g(d_i)$. (In the case of (7), we only need $\sum_{i\in A_1} g(d_{2i|1})$.) However, this quantity is not always known. When it is unknown, we need to estimate it. We follow the approach of Kwon et al. (2024). Define

$$N\alpha = \sum_{i \in U} g(d_i).$$

Assuming N is known (or that we can estimate it), we need to estimate α. Since our estimating equation is convex for any fixed α, we adopt a two-step procedure for which we estimate α and $\mathbf{w}$. In this approach we have the following loss function for the case of non-nested sampling:

$$(\hat{\alpha}, \hat{w}_2) = \arg\min_{\alpha} \arg\min_{w_2} \sum_{i \in A_2} \{G(w_{2i})\} - NK(\alpha),$$
$$\text{such that } \sum_{i \in A_2} w_{2i}\mathbf{x}_i = \hat{\mathbf{X}}_c, \text{ and } \sum_{i \in A_2} w_{2i} g(d_{2i}) = N\alpha.$$

The choice of $K(\alpha)$ is up to the analyst. We consider $K(\alpha) = \alpha$.