Data Integration with Multiple Surveys

Caleb Leedy

September 19, 2024

Overview of Data Integration

- We want to combine information from different samples.
- This is an important practical problem (Yang and Kim, 2020), (Yang, Gao, et al., 2023), (Dagdoug, Goga, and Haziza, 2023).
- We want to combine summary-level information from different sources because we do not have access to individual observations from other sources.
- We focus on data integration where all of our sources are probability samples.

Goals of Data Integration

We want to:

1. Combine information from multiple data sets,
2. In a way that is efficient, and
3. Approximately design unbiased.

Motivating Example

Sample	X_1	X_2	X_3	Y
A_0	✓	✓		✓
A_1		✓	✓	
A_2	✓		✓	
A_3	✓	✓		

Multi-Source Debiased Calibration

- This builds on the work of Kwon, Kim, and Qiu (2024).
- We construct a debiased calibration estimator within a two-phase sampling framework.

Two-Phase Sampling

- From a finite population U, one can take a two-phase sample by first selecting a Phase 1 sample from U denoted as A_1 in which one observed $(X_i)_{i=1}^{n_1}$, and then selecting a Phase 2 sample, denoted as A_2, from A_1 in which one observed $(X_i, Y_i)_{i=1}^{n_2}$.
- For additional references on two-phase sampling see Neyman (1938), Chen and Rao (2007), Legg and Fuller (2009), Hidiroglou and Särndal (1998) or specific chapters in Fuller (2009) and Kim (2024).

Existing Two-Phase Sampling Estimators

- Double Expansion Estimator
- Two-phase regression estimator

Notation

- Let π_{1i} be the probability that element i is selected into the Phase 1 sample A_1 from the finite population U.
- Let $\pi_{2i|1}$ be the conditional probability that element i is selected into the Phase 2 sample given that $i \in A_1$.
- Let $d_{1i} = 1/\pi_{1i}$ and $d_{2i|1} = 1/\pi_{2i|1}$.

Double Expansion Estimator

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

- Design unbiased,
- But does not incorporate information from Phase 1 sample

Two-Phase Regression Estimator

- To incorporate information from the Phase 1 sample, we can use a regression estimator,

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

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

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

Understanding the Two-Phase Regression Estimator

- Calibration estimator: $\hat{Y}_{\text{tp,reg}} = \sum_{i \in A_2} d_{1i} \hat{w}_{2i|1} y_i$ where

$$\begin{aligned}\hat{w}_{2i} = &\operatorname*{arg\,min}_{w_{2|1}} \sum_{i \in A_2} (w_{2i|1} - d_{2i|1})^2 q_i \\ &\text{such that} \sum_{i \in A_2} d_{1i} w_{2i|1} \mathbf{x}_i = \sum_{i \in A_1} d_{1i} \mathbf{x}_i.\end{aligned}$$

Regression as Calibration

- The regression estimator as a calibration estimator was noted by Deville and Sarndal (1992).
- They extended the calibration estimator to include other loss functions besides squared error loss.
- Their generalized loss function minimizes,

$$\sum_{i \in A_2} G(w_{2i|1}, d_{2i|1}) q_i$$

for $G(\cdot)$ that is non-negative, strictly convex with respect to $w_{2i|1}$, with a minimum at $g(d_{2i|1}, d_{2i|1})$, and defined on an interval containing $d_{2i|1}$ with $g(w_{2i|1}, d_{2i|1}) = \partial G / \partial w_{2i|1}$ continuous.

Debiased Calibration

- The debiased calibration technique comes from Kwon, Kim, and Qiu (2024).
- Instead of using a generalized loss function $G(w, d)$ like Deville and Sarndal (1992), debiased calibration uses a generalized entropy function (Gneiting and Raftery, 2007) $G(w)$ and includes a term $g(d_{2i|1})$ into the calibration.

Debiased Calibration vs. Generalized Calibration

- The motivation behind debiased calibration is that one would like to have design consistency be separated from minimizing the variance (or other loss function).
- In a regression estimator, these are separate

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

But in the generalized calibration of Deville and Sarndal (1992), these are not.

Extending Debiased Calibration

1. Two-Phase Sampling
2. Non-nested Two-Phase Sampling
3. Multi-Source Sampling

Two-Phase Debiased Calibration

- Let $\mathbf{z}_i = (\mathbf{x}_i^T / q_i, g(d_{2i|1}))^T$. The proposed two-phase debiased calibration estimator is $\hat{Y}_{\mathrm{DCE}} = \sum_{i \in A_2} w_{1i} \hat{w}_{2i|1} y_i$ where

$$\hat{w}_{2i|1} = \underset{w_{2|1}}{\arg\min} \sum_{i \in A_2} w_{1i} G(w_{2i|1}) q_i \qquad (2)$$
$$\text{such that} \sum_{i \in A_2} w_{1i} w_{2i|1} \mathbf{z}_i q_i = \sum_{i \in A_1} w_{1i} \mathbf{z}_i q_i$$

Theoretical Results: Design Consistency

Theorem (Design Consistency)

Let $\boldsymbol{\lambda}^*$ *be the probability limit of* $\hat{\boldsymbol{\lambda}}$. *Under some regularity conditions,* $\boldsymbol{\lambda}^* = (\mathbf{0}^T, 1)^T$ *and*

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

where

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

and

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

Theoretical Results

Equivalently, we have with $g_i = g(\pi_{2i|1}^{-1})q_i$,

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

for

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

Simulation Study

- Goal 1: Show that our debiased calibration estimator is indeed unbiased.
- Goal 2: Demonstrate that the proposed estimator is more efficient than the classical two-phase regression estimator.

Simulation Study

- We construct a simulation and compare: π^*-estimator, regression estimator, debiased calibration with a known population summary constraints and debiased calibration with estimated population summary constraints.

Simulation Study: Setup

For a finite population of size $N = 10,000$, and $n_1 = 1000$,

- $X_{1i} \overset{ind}{\sim} N(2, 1)$
- $X_{2i} \overset{ind}{\sim} \text{Unif}(0, 4)$
- $Z_i \overset{ind}{\sim} N(0, 1)$
- $\varepsilon_i \overset{ind}{\sim} N(0, 1)$
- $Y_i = 3X_{1i} + 2X_{2i} + 0.5Z_i + \varepsilon_i$
- $\pi_{1i} = n_1/N$
- $\pi_{2i|1} = \max(\min(\Phi_3(z_i - 1), 0.7), 0.02)$.

where Φ_3 is the CDF of a t-distribution with 3 degrees of freedom.

Simulation Study: Algorithms

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):
$$\hat{Y}_{\text{reg}} = \sum_{i \in A_1} \frac{\mathbf{x}_i' \hat{\beta}}{\pi_{1i}} + \sum_{i \in A_2} \frac{1}{\pi_{1i}\pi_{2i|1}} (y_i - \mathbf{x}_i' \hat{\beta}).$$
3. Debiased Calibration with Population Constraints (DC-Pop): This solves
$$\arg\min_{w_{2|1}} \sum_{i \in A_2} w_{1i} G(w_{2i}) \text{ such that } \sum_{i \in A_2} w_{1i} w_{2i|1} \mathbf{z}_i = \sum_{i \in U} \mathbf{z}_i.$$
4. Debiased Calibration with Estimated Population Constraints (DC-Est): This solves Equation (2) with $q_i = 1$.

Simulation Study: Results

- $B = 1000$ simulation runs.
- Let $\hat{Y}^{(b)}$ be the estimate of the bth simulation.
- Bias: $B^{-1} \sum_{b=1}^{B} \hat{Y}^{(b)} - \bar{Y}_N$
- RMSE: $\sqrt{\text{Var}_{\text{MC}}(\hat{Y} - \bar{Y}_N)}$ where $\text{Var}_{\text{MC}}(x) = \frac{1}{B-1} \sum_{b=1}^{B} (x^{(b)})^2$.
- 95% empirical confidence interval:

$$B^{-1} \sum_{b=1}^{B} I \left(|\hat{Y}^{(b)} - \bar{Y}_N| \leq \Phi(0.975) \sqrt{\hat{V}(\hat{Y}^{(b)})^{(b)}} \right)$$

- A T-test that assesses the unbiasedness of each estimator.

$$T = \frac{|\text{Bias}|}{\sqrt{\text{Var}_{\text{MC}}(\hat{Y})/B}}$$

Simulation Study: Results

Est	Bias	RMSE	EmpCI	Ttest
π^*	-0.050	0.793	0.942	1.986
TP-Reg	0.005	0.153	0.947	1.131
DC-Pop	0.002	0.092	0.968	0.677
DC-Est	0.001	0.139	0.951	0.243

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

Simulation Study: Discussion

- The debiased calibration two-phase estimator is unbiased.
- The debiased calibration two-phase estimator is more efficient than the classical regression estimator.

Non-Nested Sampling

- In non-nested two-phase sampling we have the Phase 1 sample, $A_1 = (\mathbf{X}_i)_{i=1}^{n_1}$, and the Phase 2 sample $A_2 = (\mathbf{X}_i, Y_i)_{i=1}^{n_2}$ where A_1 is independent of A_2 (Hidiroglou, 2001).
- Unlike two-phase sampling, we have two independent Horvitz-Thompson estimators of the total of $\mathbf{X}$,

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

  where $d_{2i} = \pi_{2i}^{-1}$ and π_{2i} is the first-order inclusion probability for $i \in A_2$.

Combining Information

- We can combine these estimates using the effective sample size (Kish, 1965) to get

$$\widehat{\mathbf{X}_c} = (n_{1,e}\widehat{\mathbf{X}_1} + n_{2,e}\widehat{\mathbf{X}_2})/(n_{1,e} + n_{2,e})$$

where $n_{1,e}$ and $n_{2,e}$ are the effective sample size for A_1 and A_2 respectively.

Non-Nested Regression Estimator

- We can define a regression estimator as

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

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

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

Debiased Calibration for Non-Nested Two-Phase Sampling

- The non-nested two-phase sampling debiased calibration estimator $\hat{Y}_{\mathrm{NNE}} = \sum_{i \in A_2} \hat{w}_{2i} y_i$ where

$$\hat{w}_2 = \arg\min_{w} \sum_{i \in A_2} G\left(w_{2i}\right) q_i \tag{3}$$

$$\text{with } \sum_{i \in A_2} w_{2i} \mathbf{x}_i = \widehat{\mathbf{X}_c}$$

$$\text{and } \sum_{i \in A_2} w_{2i} g(d_{2i}) q_i = \sum_{i \in U} g(d_{2i}) q_i$$

Notation

- Define

$$\widehat{\mathbf{T}} = \left[\sum_{i \in U} g(d_{2i}) q_i \right]^{\widehat{\mathbf{X}}_c}$$

Theoretical Results: Design Consistency

Theorem (Design Consistency)

Allowing $\boldsymbol{\lambda}^*$ to be the probability limit of $\hat{\boldsymbol{\lambda}}$, under some regularity conditions, $\hat{Y}_{\mathrm{NNE}} = \hat{Y}_{\ell,\mathrm{NNE}}(\boldsymbol{\lambda}^*, \boldsymbol{\phi}^*) + O_p(Nn_2^{-1})$ *where*

$$\hat{Y}_{\ell,\mathrm{NNE}}(\boldsymbol{\lambda}^*, \boldsymbol{\phi}^*) = \hat{Y}_2 + \left(\hat{\mathbf{T}} - \sum_{i \in A_2} d_{2i}\mathbf{z}_i q_i\right)\boldsymbol{\phi}^*$$

and

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

Theoretical Results: Variance Estimation

Theorem (Variance Estimation)

Under regularity conditions, and particular choice of q_i, the variance of $\hat{Y}_{\mathrm{NNE}}$ is

$$\mathit{Var}(\hat{Y}_{\mathrm{NNE}}(\hat{\lambda})) = (\boldsymbol{\phi}_1^*)^T \mathit{Var}(\widehat{\mathbf{X}}_c)\boldsymbol{\phi}_1^* + \sum_{i \in U}\sum_{j \in U} \frac{\Delta_{2ij}}{\pi_{2i}\pi_{2j}}(y_i - \mathbf{z}_i\boldsymbol{\phi}^* q_i)(y_j - \mathbf{z}_j\boldsymbol{\phi}^* q_j)$$

Theoretical Results: Variance Estimation

We can estimate the variance using

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

where

$$\hat{\boldsymbol{\phi}} = \begin{bmatrix} \hat{\boldsymbol{\phi}}_1 \\ \hat{\phi}_2 \end{bmatrix} = \left(\sum_{i \in A_2} \frac{\mathbf{z}_i \mathbf{z}_i^T q_i}{g'(d_{2i})} \right)^{-1} \sum_{i \in A_2} \frac{\mathbf{z}_i y_i}{g'(d_{2i})}$$

Simulation Study

- We want to make sure that we can incorporate additional sampling information into the non-nested design, and that we are asymptotically equivalent to a regression estimator.

Simulation Study: Setup

- $X_{1i} \stackrel{ind}{\sim} N(2,1)$
- $X_{2i} \stackrel{ind}{\sim} \text{Unif}(0,4)$
- $Z_i \stackrel{ind}{\sim} N(0,1)$
- $\varepsilon_i \stackrel{ind}{\sim} N(0,1)$
- $Y_i = 3X_{1i} + 2X_{2i} + z_i + \varepsilon_i$
- $\pi_{1i} = n_1/N$
- $\pi_{2i} = \max(\min(\Phi_3(z_i - 2.5), 0.9), 0.01)$

where Φ_3 is the CDF of a t-distribution with 3 degrees of freedom.

Simulation Study: Setup

- $N = 10,000$
- $n_1 = 1000$
- $E[n_2] \approx 725$
- $B = 1000$

Simulation Study: Algorithms

1. HT-estimator: $\hat{Y}_2 = N^{-1} \sum_{i \in A_2} d_{2i} y_i$,

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

Simulation Study: Algorithms

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

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

4. Debiased Calibration with Estimated Population Constraints (DC-Est): This solves Equation (3) with $q_i = 1$.

Simulation Study: Results

Est	Bias	RMSE	EmpCI	Ttest
HT	-0.023	0.539	0.941	1.365
Reg	-0.003	0.128	0.970	0.765
DC-Pop	0.003	0.068	0.929	1.492
DC-Est	-0.003	0.125	0.966	0.668

Table: 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.

Simulation Study: Discussion

- The debiased calibration non-nested two-phase estimator is slightly more efficient than the regression estimator.

Notation

- Assume that for A_0 we observe $(X^{(0)}, Y)_{i=1}^{n_0}$.
- For $A_1, \ldots, A_M$, we observe $(X^{(m)})_{i=1}^{n_m}$

Combining Information via GLS

- For each A_m, we construct Horvitz-Thompson estimates of the mean each of the observed $\mathbf{X}^{(m)}$ variables, and combine them with a GLS estimate.
- For example, if we have the following setup:

Sample	X_1	X_2	X_3	Y
A_0	✓	✓		✓
A_1		✓	✓	
A_2	✓		✓	
A_3	✓	✓		

Combining Information via GLS

$$\underbrace{\begin{bmatrix} \hat{X}_1^{(0)} \\ \hat{X}_2^{(0)} \\ \hat{X}_2^{(1)} \\ \hat{X}_3^{(1)} \\ \hat{X}_1^{(2)} \\ \hat{X}_3^{(2)} \\ \hat{X}_1^{(3)} \\ \hat{X}_2^{(3)} \end{bmatrix}}_{\hat{\mathbf{X}}} = \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}}_{\mathbf{D}} \underbrace{\begin{bmatrix} \mu_{X_1} \\ \mu_{X_2} \\ \mu_{X_3} \end{bmatrix}}_{\boldsymbol{\mu}} + \mathbf{e}$$

where $\mathbf{e} \sim N(\mathbf{0}, \mathbf{V})$ and $\mathbf{V} = \text{diag}(\mathbf{V}_0, \mathbf{V}_1, \mathbf{V}_2, \mathbf{V}_3)$ with

Combining Information via GLS

$$\mathbf{V}_0 = \begin{bmatrix} V_{X_1}^{(0)} & C_{X_1,X_2}^{(0)} \\ C_{X_1,X_2}^{(0)} & V_{X_2}^{(0)} \end{bmatrix}, \mathbf{V}_1 = \begin{bmatrix} V_{X_2}^{(1)} & C_{X_2,X_3}^{(1)} \\ C_{X_2,X_3}^{(1)} & V_{X_3}^{(1)} \end{bmatrix},$$

$$\mathbf{V}_2 = \begin{bmatrix} V_{X_1}^{(2)} & C_{X_1,X_3}^{(2)} \\ C_{X_1,X_3}^{(2)} & V_{X_3}^{(2)} \end{bmatrix}, \mathbf{V}_3 = \begin{bmatrix} V_{X_1}^{(3)} & C_{X_1,X_2}^{(3)} \\ C_{X_1,X_2}^{(3)} & V_{X_2}^{(3)} \end{bmatrix}.$$

Combining Information via GLS

- Then we have the GLS estimate of $\bar{X}_N$ is

$$\hat{\mathbf{X}}_{\mathrm{GLS}} = (\mathbf{D}^T\mathbf{V}^{-1}\mathbf{D})\mathbf{D}^T\mathbf{V}^{-1}\hat{\mathbf{X}}.$$

- We use the estimate $\hat{\mathbf{X}}_{\mathrm{GLS}}$ as a constraint in our debiased calibration model.

Debiased Calibration with Multiple Sources

The multi-source debiased calibration estimator is then $\hat{Y}_{\text{MS}} = \sum_{i \in A_0} \hat{w}_{0i} y_i$ where

$$\hat{w}_0 = \arg\min_{w} \sum_{i \in A_0} G(w_i) q_i$$

$$\text{such that} \sum_{i \in A_0} w_i x_i^{(0)} = N \hat{\mathbf{X}}_{\text{GLS}}^{(0)}$$

$$\text{and} \sum_{i \in A_0} w_i g(d_{0i}) q_i = \sum_{i \in U} g(d_{0i}) q_i.$$

Debiased Calibration with Multiple Sources

- If we let $\hat{\mathbf{T}} = \left((\hat{\mathbf{X}}^{(0)}_{\mathrm{GLS}})^T, \sum_{i \in U} g(d_{0i}) q_i \right)^T$ and $\mathbf{z}_i = ((\mathbf{x}_i^{(0)})^T / q_i, g(d_{0i}))^T$, then we can solve for $\hat{w}_0$ using the Lagrange multiplier approach of

$$\hat{w}_0 = \arg\min_w \sum_{i \in A_0} G(w_i) q_i - \lambda \left(\hat{\mathbf{T}} - \sum_{i \in A_0} w_i \mathbf{z}_i q_i \right). \tag{4}$$

Theoretical Results: Design Consistency

Theorem (Design Consistency)

Let $\boldsymbol{\lambda}^*$ be the probability limit of $\hat{\boldsymbol{\lambda}}$. Under regularity conditions,

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

where

$$\hat{Y}_{\ell}(\boldsymbol{\lambda}^*, \boldsymbol{\phi}^*) = \sum_{i \in A_0} d_{0i} y_i + \left(\hat{\mathbf{T}} - \sum_{i \in A_0} d_{0i} \mathbf{z}_i q_i \right) \boldsymbol{\phi}^* \tag{5}$$

and

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

Theoretical Results: Variance Estimation

Theorem (Variance Estimation)

Under regularity conditions,

$$V(\hat{Y}_{\mathrm{MS}}) = (\boldsymbol{\phi}_1^{(0)*})^T \mathit{Var}(\hat{\mathbf{X}}_{\mathrm{GLS}}^{(0)})(\boldsymbol{\phi}_1^{(0)*})^T + \sum_{i \in U} \sum_{j \in U} \frac{\Delta_{0ij}}{\pi_{0i}\pi_{0j}} (y_i - \mathbf{z}_i^{(0)} \boldsymbol{\phi}^{(0)*} q_i)(y_j - \mathbf{z}_j^{(0)} \boldsymbol{\phi}^{(0)*} q_j).$$

We can estimate the variance with

$$\hat{V}(\hat{Y}_{\mathrm{MS}}) = (\hat{\boldsymbol{\phi}}_1{}^{(0)})^T \widehat{\mathit{Var}}(\hat{\mathbf{X}}_{\mathrm{GLS}}^{(0)})(\hat{\boldsymbol{\phi}}_1{}^{(0)})^T + \sum_{i \in A_0} \sum_{j \in A_0} \frac{\Delta_{0ij}}{\pi_{0ij}\pi_{0i}\pi_{0j}} (y_i - \mathbf{z}_i^{(0)} \hat{\boldsymbol{\phi}}^{(0)} q_i)(y_j - \mathbf{z}_j^{(0)} \hat{\boldsymbol{\phi}}^{(0)} q_j).$$

Simulation

- We want to check if we can successfully incorporate multiple samples.
- We want to do no worse than a regression estimator.

Simulation: Setup

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

Simulation: Setup

We observe the following columns in each sample

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

Simulation: Estimators

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 (3) with only using information from Samples A_0 and A_1,
3. Multi-Source proposed (MSEst): This is the proposed estimator from Equation (4).
4. Multi-Source population (MSPop): This is the proposed estimator from Equation (4) with using the true value of $\hat{\mathbf{T}}$ from the population,

Simulation: Results

Est	Bias	RMSE	EmpCI	Ttest
HT	-0.0062	0.5687	0.947	0.3422
NNReg	-0.0016	0.1065	0.974	0.4876
MSPop	0.0002	0.0681	0.945	0.0728
MSEst	-0.0019	0.0967	0.938	0.6300

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

Discussion

- We have extended the debiased calibration of Kwon, Kim, and Qiu (2024) to the two-phase, non-nested two-phase, and multi-source setting.
- It appears to work well and it is more efficient than a regression estimator.
- We have theory for the design consistency and variance estimation.

Thank You!

References I

- Chen, Jiahua and JNK Rao (2007). "Asymptotic normality under two-phase sampling designs". In: *Statistica sinica*, pp. 1047–1064.
- Dagdoug, Mehdi, Camelia Goga, and David Haziza (2023). "Model-assisted estimation through random forests in finite population sampling". In: *Journal of the American Statistical Association* 118(542), pp. 1234–1251.
- Deville, Jean-Claude and Carl-Erik Sarndal (1992). "Calibration estimators in survey sampling". In: *Journal of the American statistical Association* 87(418), pp. 376–382.
- Fuller, Wayne A (2009). *Sampling statistics*. John Wiley & Sons.
- Gneiting, Tilmann and Adrian E Raftery (2007). "Strictly proper scoring rules, prediction, and estimation". In: *Journal of the American statistical Association* 102(477), pp. 359–378.

References II

Hidiroglou, MA (2001). “Double sampling”. In: *Survey methodology* 27(2), pp. 143–154.

Hidiroglou, MA and CE Särndal (1998). “Use of auxiliary information for two-phase sampling”. In: *Survey Methodology* 24, pp. 11–20.

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

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

Kott, P Stukel and DM Stukel (1997). “Can the jackknife be used with a two-phase sample”. In: *Survey Methodology* 23(2), pp. 81–89.

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

References III

Legg, Jason C and Wayne A Fuller (2009). "Two-phase sampling". In: *Handbook of statistics*. Vol. 29. Elsevier, pp. 55–70.

Neyman, Jerzy (1938). "Contribution to the theory of sampling human populations". In: *Journal of the American Statistical Association* 33(201), pp. 101–116.

Randles, Ronald H (1982). "On the asymptotic normality of statistics with estimated parameters". In: *The Annals of Statistics*, pp. 462–474.

Yang, Shu, Chenyin Gao, et al. (2023). "Elastic integrative analysis of randomised trial and real-world data for treatment heterogeneity estimation". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 85(3), pp. 575–596.

Yang, Shu and Jae Kwang Kim (2020). "Statistical data integration in survey sampling: A review". In: *Japanese Journal of Statistics and Data Science* 3, pp. 625–650.

Proof

- The first order conditions for Equation (2) show that

$$g(w_{2i|1})w_{1i}q_i - w_{1i}\boldsymbol{\lambda}^T\mathbf{z}_iq_i = 0.$$

- Hence, $\hat{w}_{2i}(\boldsymbol{\lambda}) = g^{-1}(\boldsymbol{\lambda}^T\mathbf{z}_i)$ and $\hat{\boldsymbol{\lambda}}$ is determined by Equation (6).

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

- When the sample size gets large, we have $\hat{w}_{2i|1}(\hat{\boldsymbol{\lambda}}) \to d_{2i|1}$
- Then $\hat{\boldsymbol{\lambda}} \to \boldsymbol{\lambda}^*$ where $\boldsymbol{\lambda}^* = (\mathbf{0}^T, 1)^T$.

Proof (continued)

- Using the linearization technique of Randles (1982), we can construct a regression estimator,

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

- We choose ϕ^* such that $E\left[\frac{\partial}{\partial \boldsymbol{\lambda}} \hat{Y}_\ell(\boldsymbol{\lambda}^*, \boldsymbol{\phi}^*)\right] = 0$. Since $g^{-1}(\boldsymbol{\lambda}^* \mathbf{z}_i) = g^{-1}(g(d_{2i|1})) = d_{2i|1}$ and $(g^{-1})'(x) = 1/g'(g^{-1}(x))$,

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

Proof (continued)

- The linearization estimator is

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

- Using a Taylor expansion yields,

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

Proof

- The biggest modification for this proof is that the total for $\mathbf{X}$ is estimated from both samples using $\widehat{\mathbf{X}_c}$ instead of $\widehat{\mathbf{X}}_{\mathrm{HT}}$ from the Phase 1 sample.
- Since $\hat{Y}_{\mathrm{NNE}} = \sum_{i \in A_2} \hat{w}_{2i}(\hat{\boldsymbol{\lambda}}) y_i$, to linearize using the linearization technique of Randles (1982), we have

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

Proof (continued)

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

$$\boldsymbol{\phi}^* = \begin{bmatrix} \phi_1^* \\ \phi_2^* \end{bmatrix} = \left(\sum_{i \in U} \frac{\pi_{2i} \mathbf{z}_i \mathbf{z}_i^T q_i}{g'(d_{2i})}\right)^{-1} \sum_{i \in U} \frac{\pi_{2i} \mathbf{z}_i y_i}{g'(d_{2i})}.$$

- By a Taylor expansion around $\hat{\boldsymbol{\lambda}}$,

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

Proof

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

Proof (continued)

Since $\hat{\mathbf{X}}_c = W\hat{\mathbf{X}}_1 + (1-W)\hat{\mathbf{X}}_2$.

$$= (\phi_1^*)^T \text{Var}(\hat{\mathbf{X}}_c)\phi_1^* + \sum_{i \in U}\sum_{j \in U} \frac{\Delta_{2ij}}{\pi_{2i}\pi_{2j}}(y_i - \mathbf{z}_i \boldsymbol{\phi}^* q_i)(y_j - \mathbf{z}_j \boldsymbol{\phi}^* q_j)$$

$$+ 2(1-W)\phi_1^* \sum_{i \in U}\sum_{j \in U} \Delta_{2ij} \frac{x_i}{\pi_{2i}} \frac{(y_j - \mathbf{z}_j \phi_1^* q_j)}{\pi_{2j}}$$

and the covariance term is $O(1)$ by the choice of q_i.