Data Integration with Multiple Surveys

Caleb Leedy and Jae Kwang Kim

April 4, 2024

Variance with Estimated Population Constraints

Setup

This summarizes the setup described in `main.tex`. Suppose that we have $K = 3$ surveys where we observe different variables:

- A_1: x_1, x_2, x_3, y_1
- A_2: $x_1, \quad x_3, y_2$
- A_3: $x_1, x_2, \quad y_3$

We assume that these surveys are independent, and our estimation procedure consists of two steps.

Step 1: GLS Estimation

Let $\theta = (\mu_1, \mu_2, \mu_3)' = (E[x_1], E[x_2], E[x_3])'$. We first estimate θ using the following GLS

$$\begin{bmatrix} \hat{\mu}_{1,1} \\ \hat{\mu}_{1,2} \\ \hat{\mu}_{1,3} \\ \hat{\mu}_{2,1} \\ \hat{\mu}_{2,3} \\ \hat{\mu}_{3,1} \\ \hat{\mu}_{3,2} \end{bmatrix} := \underbrace{\begin{bmatrix} n_1^{-1} \sum_{i \in A_1} x_{1i} \\ n_1^{-1} \sum_{i \in A_1} x_{2i} \\ n_1^{-1} \sum_{i \in A_1} x_{3i} \\ n_2^{-1} \sum_{i \in A_2} x_{1i} \\ n_2^{-1} \sum_{i \in A_2} x_{3i} \\ n_3^{-1} \sum_{i \in A_3} x_{1i} \\ n_3^{-1} \sum_{i \in A_3} x_{2i} \end{bmatrix}}_{\hat{\theta}} = \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}}_{X} \underbrace{\begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix}}_{\theta} + \mathbf{e}$$

where $\mathbf{e} \sim (0, V)$ and

$$V = \begin{bmatrix} V_1 & 0_{3\times 2} & 0_{3\times 2} \\ 0_{2\times 3} & V_2 & 0_{2\times 2} \\ 0_{2\times 3} & 0_{2\times 2} & V_3 \end{bmatrix}$$

and V_1, V_2, and V_3 are known. Then the GLS estimator is

$$\hat{\theta}_{GLS} = (X'V^{-1}X)^{-1}X'V^{-1}\hat{\theta}.$$

Step 2: Debiased Calibration

Let $H_k(X)$ output all of the observed vectors of X for sample k. Then using the same notation as Kwon, Kim, and Qiu 2024, the optimal weights $\hat{\mathbf{w}}^{(k)}$ for a sample A_k solve

$$\begin{aligned}\hat{\mathbf{w}}^{(k)} =& \text{argmin}_{\mathbf{w}} \sum_{i \in A_k} G(w_i) \\ &\text{such that } \sum_{i \in A_k} w_i^{(k)} H_k(x_i) = H_k(\hat{\theta}_{GLS})N \text{ and } \sum_{i \in A_k} w_i^{(k)} g(d_i) = \sum_{i \in U} g(d_i).\end{aligned} \tag{1}$$

Result

Before deriving the variance, I reintroduce some of the notation from Kwon, Kim, and Qiu 2024. Let $G : \mathcal{V} \to \mathbb{R}$ be a strictly convex and differentiable function with derivative $g(w) = dG(w)/dw$ and define the convex conjugate function of $G(w)$ as $F(w) = -G(g^{-1}(w)) + g^{-1}(w)w$ with a derivative $f(w) = dF(w)/dw$. Note, that $f(w) = g^{-1}(w)$. The primal problem of Equation 1 has a dual formulation of

$$\hat{\lambda} = \text{argmin}_{\lambda \in \Lambda_{A_k}} \sum_{i \in A_k} F(\lambda^T \mathbf{z}_i^{(k)}) - \lambda^T \sum_{i \in U} \mathbf{z}_i^{(k)}.$$

In this case $z^{(k)} = (H_k(X), g(d_i))$, $d_i = \pi_i^{-1}$, $\hat{\lambda} = (\hat{\lambda}_1^T, \hat{\lambda}_2)^T$, and $\Lambda_A = \{\lambda : \lambda^T z_i^{(k)} \in g(\mathcal{V}) \text{ for all } i \in A\}$. We still define λ as having dimension $p+1$, but p is now the dimension of $H_k(X)$. The parameter λ is also a function of the sample A_k and so should be $\lambda = \lambda^{(k)}$, but we suppress this notation for convenience. Solutions to the primal and dual problem satisfy

$$\hat{w}_i^{(k)} = \hat{w}_i^{(k)}(\hat{\lambda}) = f(\hat{\lambda}^T z_i) = g^{-1}(\hat{\lambda}_1^T H_k(x_i) + \hat{\lambda}_2 g(d_i)).$$

For now, I am only going to derive the variance for the estimate $\hat{Y}_1 = \sum_{i \in A_1} \hat{w}_i^{(1)} y_{1i}$. In this case we have $H_1(X) = X$, however, other samples can be derived in a similar manner. We define that matrix

$$\begin{bmatrix} \Sigma_{xx} & \Sigma_{xg} \\ \Sigma_{gx} & \Sigma_{gg} \end{bmatrix} = N^{-1} \sum_{i \in U} \frac{\pi_i}{g'(d_i)} \begin{bmatrix} x_i x_i^T & g(d_i) x_i^T \\ x_i g(d_i) & g(d_i)^2 \end{bmatrix}.$$

Now we can find the derivative of $\hat{\Lambda}$ with respect to θ.

Lemma 1.

$$\frac{\partial \hat{\lambda}(\theta_N)}{\partial \theta_N} = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xg} \\ \Sigma_{gx} & \Sigma_{gg} \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

Proof. This follows the approach of Kwon, Kim, and Qiu 2024 in Appendix A.3. Notice that taking a derivative with respect to θ in the constraints yields

$$\left[N^{-1}\sum_{i\in A_1} f'(\hat{\lambda}(\theta)^T \mathbf{z}_i^{(1)})(z_i^{(1)})^T\right]\begin{bmatrix}\hat{\lambda}_1'(\theta)\\ \hat{\lambda}_2'(\theta)\end{bmatrix} = \begin{bmatrix}1\\0\end{bmatrix}.$$

This means that

$$\begin{bmatrix}\hat{\lambda}_1'(\theta_N)\\ \hat{\lambda}_2'(\theta_N)\end{bmatrix} = \left[N^{-1}\sum_{i\in A_1} f'(\hat{\lambda}(\theta_N)^T \mathbf{z}_i^{(1)})(z_i^{(1)})^T\right]^{-1}\begin{bmatrix}1\\0\end{bmatrix} \to \begin{bmatrix}\Sigma_{xx} & \Sigma_{xg}\\ \Sigma_{gx} & \Sigma_{gg}\end{bmatrix}^{-1}\begin{bmatrix}1\\0\end{bmatrix}.$$

$\square$

Since the GLS estimator satisfies $\hat{\theta}_{GLS} - \theta_N = o_p(n^{-1/2})$, we have

$$\begin{aligned}
&N^{-1}\sum_{i\in A_1}\hat{w}_i^{(1)}(\hat{\theta}_{GLS})y_{1i}\\
&= N^{-1}\sum_{i\in A_1}\hat{w}_i^{(1)}(\theta_N)y_{1i} + \left(N^{-1}\sum_{i\in A_1}\frac{\partial w_i^{(k)}(\theta_N)}{\partial\theta}y_{1i}\right)(\hat{\theta}_{GLS}-\theta_N) + o_p(n^{-1/2})\\
&= \hat{\theta}_{DC} + \left(N^{-1}\sum_{i\in A_1} f'(\hat{\lambda}^T(\theta_N)z_i^{(1)})\frac{\partial\lambda(\theta_N)^T}{\partial\theta}z_i^{(1)}y_{1i}\right)(\hat{\theta}_{GLS}-\theta_N) + o_p(n^{-1/2})\\
&= \hat{\theta}_{DC} + \gamma_{N,1:p}'(\hat{\theta}_{GLS}-\theta_N) + o_p(n^{-1/2})
\end{aligned}$$

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

$$\gamma_N = \left[\sum_{i\in U}\frac{\pi_i z_i^{(1)}(z_i^{(1)})^T}{g'(d_i)}\right]^{-1}\sum_{i\in U}\frac{\pi_i z_i^{(1)}y_{1i}}{g'(d_i)}$$

and $\gamma_{N,1:p}$ is the first p elements of γ_N. This means that the variance of the estimator of Y_1 is

$$\begin{aligned}
\text{Var}(\hat{Y}_1) &= \text{Var}\left(\sum_{i\in A_1} w_i^{(1)y_{1i}}\right)\\
&= \text{Var}(\hat{\theta}_{DC}) + \gamma_{N,1:p}'(X'V^{-1}X)^{-1}\gamma_{n,1:p} + 2\gamma_{N,1:p}'\text{Cov}(\hat{\theta}_{DC}, \hat{\theta}_{GLS}-\theta_N).
\end{aligned}$$

Dr. Kim, I think this is the point of departure from your proposed idea and what I do. I probably could use a linearization to estimate the variance but instead I do it directly. Please let me know what you think. Finally,

$$
\begin{aligned}
\mathrm{Cov}(\hat{\theta}_{DC}, \hat{\theta}_{GLS} - \theta_N) &= \mathrm{Cov}(\hat{\theta}_{DC}, \hat{\theta}_{GLS}) \\
&= (X'V^{-1}X)^{-1}X'V^{-1}n_1^{-1}\sum_{i\in U}\sum_{j\in U}\mathrm{Cov}\left(\frac{\delta_{1i}y_{1i}}{g(\hat{\lambda}_1 H_1(x_i) + \hat{\lambda}_2 g(d_i))}, \delta_{1j}H_1(x_j)\right) \\
&= (X'V^{-1}X)^{-1}X'V^{-1}n_1^{-1}\sum_{i\in U}\mathrm{Cov}\left(\frac{\delta_{1i}y_{1i}}{g(\hat{\lambda}_1 H_1(x_i) + \hat{\lambda}_2 g(d_i))}, \delta_{1i}H_1(x_i)\right) \\
&= (X'V^{-1}X)^{-1}X'V^{-1}n_1^{-1}\sum_{i\in U}\frac{H_1(x_i)y_{1i}}{g(\hat{\lambda}_1 H_1(x_i) + \hat{\lambda}_2 g(d_i))}\mathrm{Cov}\left(\delta_{1i}, \delta_{1i}\right) \\
&= (X'V^{-1}X)^{-1}X'V^{-1}n_1^{-1}\sum_{i\in U}\frac{H_1(x_i)y_{1i}}{g(\hat{\lambda}_1 H_1(x_i) + \hat{\lambda}_2 g(d_i))}(\pi_{1i}(1-\pi_{1i}))
\end{aligned}
$$

If we note that $\pi_{1i}g^{-1}(\hat{\lambda}_1 H_1(x_i) + \hat{\lambda}_2 g(d_i)) \to 1$ as $N \to \infty$ then we have

$$
\mathrm{Cov}(\hat{\theta}_{DC}, \hat{\theta}_{GLS} - \theta_N) = (X'V^{-1}X)^{-1}X'V^{-1}n_1^{-1}\sum_{i\in U}H_1(x_i)y_{1i}((1-\pi_{1i})).
$$