Survey Data Integration: Part 3

Jae-Kwang Kim

Iowa State University

October 10th, 2023
Center for Statistical Science @ Peking University

Outline

1. Introduction
2. Quasi-randomization approach
3. Outcome regression model approach
4. Examples
5. Doubly robust estimation
6. Conclusion

A motivating example: PRC data

- Pew Research Center (PRC) data in 2015: a non-probability sample data of size $n = 9,301$ with 56 variables, provided by eight different vendors with unknown sampling and data collection strategies.
- The PRC dataset aims to study the relation between people and community. We choose 9 variables, among them 8 are binary and 1 is continuous, as response variables in our analysis.
- We consider two probability samples with common auxiliary variables. The first is the Behavioral Risk Factor Surveillance System (BRFSS) survey data and the second is the Volunteer Supplement survey data from the Current Population Survey (CPS), both collected in 2015.

Comparison of covariates from three datasets

Table: Estimated Population Mean of Covariates from the Three Samples

		$\hat{X}_{PRC}$	$\hat{X}_{BRFSS}$	$\hat{X}_{CPS}$
Age category	<30	0.183	0.209	0.212
	>=30,<50	0.326	0.333	0.336
	>=50,<70	0.387	0.327	0.326
	>=70	0.104	0.131	0.126
Gender	Female	0.544	0.513	0.518
Race	White only	0.823	0.750	0.786
Race	Black only	0.088	0.126	0.125
Origin	Hispanic/Latino	0.093	0.165	0.156
Region	Northeast	0.200	0.177	0.180
Region	South	0.275	0.383	0.373
Region	West	0.299	0.232	0.235
Marital status	Married	0.503	0.508	0.528
Employment	Working	0.521	0.566	0.589
Employment	Retired	[illegible]	[illegible]	[illegible]

Comparison of covariates from three datasets (Cont'd)

		$\hat{X}_{PRC}$	$\hat{X}_{BRFSS}$	$\hat{X}_{CPS}$
Education	High school or less	0.216	0.427	0.407
Education	Bachelor's degree and above	0.416	0.263	0.309
Education	Bachelor's degree	0.221	NA	0.198
Education	Postgraduate	0.195	NA	0.111
Household	Presence of child in household	0.289	0.368	NA
Household	Home ownership	0.654	0.672	NA
Health	Smoke everyday	0.157	0.115	NA
Health	Smoke never	0.798	0.833	NA
Financial status	No money to see doctors	0.207	0.133	NA
Financial status	Having medical insurance	0.891	0.878	NA
Financial status	Household income $<$ 20K	0.161	NA	0.153
Financial status	Household income $>$100K	0.199	NA	0.233
Volunteer works	Volunteered	0.510	NA	0.248

Remark

- There are noticeable differences between the naive estimates from the PRC sample and the estimates from the two probability samples for covariates such as Origin (Hispanic/Latino), Education (High school or less), Household (with children), Health (Smoking) and Volunteer works.
- It is a strong evidence that the PRC dataset is not a representative sample for the population.

Basic Setup

- A finite population $U = \{1, \cdots, N\}$; associated unit i are
  - study variable y_i: often expensive to measure.
  - auxiliary variable $\boldsymbol{x}_i$: often correlated with y_i

Table: Data Structure for Two Samples

Sample	Type	X	Y	Sampling Weight
A	Probability Sample	✓		✓
B	Non-probability Sample	✓	✓	

- Two independent samples from the finite population U.
  1. Probability sample: observe $\boldsymbol{x}_i$ for $i \in A \subset U$.
  2. Non-Probability sample: observe $(\boldsymbol{x}_i, y_i)$ for $i \in B \subset U$.
- We wish to combine the two data sources to estimate $\theta_N = N^{-1} \sum_{i=1}^{N} y_i$.

Notation

- $w_i^{(A)}$: Sampling weight for the probability sample A
- δ_i: Sampling indicator for the sample B

$$\delta_i = \begin{cases} 1 & \text{if } i \in B \\ 0 & \text{otherwise.} \end{cases}$$

Goal

- Wish to construct propensity score (PS) weights ω_i in sample B such that we can estimate θ_N by

$$\hat{\theta}_{\mathrm{PS}} = \frac{1}{N} \sum_{i \in B} \omega_i y_i \tag{1}$$

  The PS weights should be nonnegative. That is,

$$\omega_i \geq 0, \quad \forall i \in B.$$

- Two approaches for obtaining ω_i:
  1. Pseudo-randomization approach: Use a model for δ.
     - Assume a model for sampling mechanism for B, say $\pi_i = \mathbb{P}(\delta_i = 1 \mid \boldsymbol{x}_i)$, and use $\omega_i = 1/\hat{\pi}_i$.
  2. Outcome regression model approach: Use a model for Y.

Pseudo-randomization approach (Chen et al., 2020)

1. Assume a parametric model for $\mathbb{P}(\delta_i = 1 \mid \boldsymbol{x}_i)$, say $\pi(\boldsymbol{x}_i; \phi)$.
2. If $A = U$, then we would solve

$$S(\phi) = \sum_{i=1}^{N} \{\delta_i - \pi(\boldsymbol{x}_i; \phi)\} h(\boldsymbol{x}_i; \phi) = 0$$

to estimate ϕ, where

$$h(\boldsymbol{x}_i; \phi) = \frac{1}{\pi(\boldsymbol{x}_i; \boldsymbol{\phi})\{1 - \pi(\boldsymbol{x}_i; \boldsymbol{\phi})\}} \frac{\partial}{\partial \boldsymbol{\phi}} \pi(\boldsymbol{x}_i; \boldsymbol{\phi}).$$

3. Thus, parameter ϕ is estimated by solving

$$\hat{S}\left(\boldsymbol{\phi}\right) := \sum_{i \in B} h(\boldsymbol{x}_i; \boldsymbol{\phi}) - \sum_{i \in A} w_i^{(A)} \pi(\boldsymbol{x}_i; \boldsymbol{\phi}) h(\boldsymbol{x}_i; \boldsymbol{\phi}) = 0.$$

4. Use the inverse of $\hat{\pi}_i = \pi(\boldsymbol{x}_i; \hat{\boldsymbol{\phi}})$ as the propensity weight for sample B.

Conditional ML approach (Kim and Kwon, 2024)

- We are interested in estimating ϕ for $\pi(\boldsymbol{x}_i; \phi)$ using the observations in the combined sample $A \cup B$.
- Derive the conditional inclusion probability

$$\begin{aligned} P(i \in B \mid i \in A \cup B) &= \frac{\pi_i^{(B)}(\phi)}{\pi_i^{(A)} + \pi_i^{(B)}(\phi) - \pi_i^{(A)} \cdot \pi_i^{(B)}(\phi)} \\ &:= \pi_{i,\mathrm{cond}}(\phi), \end{aligned}$$

where $\pi_i^{(B)}(\phi) = \pi(\boldsymbol{x}_i; \phi)$ and $\pi_i^{(A)} = \{w_i^{(A)}\}^{-1}$ is the first-order inclusion probability of sample A.

- We can estimate ϕ by maximizing the conditional log-likelihood

$$\ell_{\text{cond}}(\phi) = \sum_{i \in A \cup B} \left[\delta_i \log \pi_{i,\text{cond}}(\phi) + (1 - \delta_i) \log\{1 - \pi_{i,\text{cond}}(\phi)\} \right]. \tag{2}$$

- In practice, the first-order inclusion probabilities $\pi_i^{(A)}$ are unknown outside sample A. In this case, (2) cannot be used directly. One way to handle this problem is to compute

$$\tilde{\pi}_i^{(A)} = P(I_i^{(A)} = 1 \mid \boldsymbol{x}_i)$$

which can be obtained by $\tilde{\pi}_i^{(A)} = 1/\tilde{w}_i^{(A)}$ where

$$\tilde{w}_i^{(A)} = E\{w_i^{(A)} \mid \boldsymbol{x}_i, I_i^{(A)} = 1\} \tag{3}$$

following the result of Pfeffermann and Sverchkov (1999).

- Use $\tilde{\pi}_i^{(A)} = 1/\tilde{w}_i^{(A)}$ in place of $\pi_i^{(A)}$ in the conditional likelihood.

3. Outcome Regression model approach

- Recall two approaches for obtaining ω_i:
  1. Pseudo-randomization approach: Use a model for δ.
  2. Outcome regression model approach: Use a model for Y.
- The pseudo-randomization approach gives a design-based flavor but it is still model-based. However, the subject matter knowledge to build a model for δ is not available, in general.
- On the other hand, the subject-matter knowledge to build a model for Y can be obtained from other surveys (or from domain experts).
- Therefore, if the subject-matter knowledge is available for the study outcome, the outcome regression model approach is more attractive in developing propensity score weights for data integration.

Assumptions

1. Ignorability (MAR):

$$f(y \mid \boldsymbol{x}) = f(y \mid \boldsymbol{x}, \delta = 1)$$

   where

$$\delta_i = \begin{cases} 1 & \text{if } i \in B \\ 0 & \text{otherwise.} \end{cases}$$

2. Positivity

$$P(\delta = 1 \mid \boldsymbol{x}) > \epsilon > 0$$

   almost everywhere.

Motivation

- Assume the outcome regression model

$$Y_i = \boldsymbol{x}_i' \boldsymbol{\beta} + e_i$$

  where $e_i \mid \boldsymbol{x}_i \sim (0, v_i)$, with $v_i = v(\boldsymbol{x}_i)$ known and $\boldsymbol{x}_i$ includes an intercept term.

- Writing $\theta_N = N^{-1} \sum_{i=1}^{N} y_i$, we can express

$$\hat{\theta}_{\mathrm{PS}} - \theta_N = N^{-1} \left(\sum_{i \in B} \omega_i \boldsymbol{x}_i - \sum_{i=1}^{N} \boldsymbol{x}_i \right)' \boldsymbol{\beta} + N^{-1} \sum_{i \in B} \omega_i e_i - N^{-1} \sum_{i=1}^{N} e_i$$

- The second term has zero expectation. Thus, the PS estimator $\hat{\theta}_{\mathrm{PS}}$ is unbiased if

$$\sum_{i \in B} \omega_i \boldsymbol{x}_i = \sum_{i=1}^{N} \boldsymbol{x}_i \tag{4}$$

  holds. Condition (4) is often called the covariat-balancing condition.

- Under (4), the variance of $\hat{\theta}_{\rm PS}$ is

$$V\left(\hat{\theta}_{\rm PS} - \hat{\theta}_N\right) = N^{-2}\sum_{i\in B}\omega_i^2 v_i - 2N^{-2}\sum_{i\in B}\omega_i v_i + N^{-2}\sum_{i\in U} v_i.$$

- Thus, if we expand $\mathbf{x}_i$ to include v_i, we have

$$V\left(\hat{\theta}_{\rm PS} - \hat{\theta}_N\right) = N^{-2}\sum_{i\in B}\omega_i^2 v_i - N^{-2}\sum_{i\in U} v_i.$$

  we can find the minimizer of

$$Q(\omega) = \sum_{i\in B}\omega_i^2 v_i$$

  subject to (4).

- Finding weights satisfying some covariate-balancing constraint in (4) is often called the calibration weighting problem in survey sampling.

- Using Lagrange multiplier method, the solution is

$$\hat{\omega}_i = \hat{\boldsymbol{\lambda}}' \boldsymbol{x}_i / v_i$$

  where $\hat{\boldsymbol{\lambda}}$ satisfies (4).

- The calibration equation for $\boldsymbol{\lambda}$ is

$$\sum_{i \in B} \hat{\omega}_i \boldsymbol{x}_i' = \sum_{i \in B} \left(\hat{\boldsymbol{\lambda}}' \boldsymbol{x}_i / v_i \right) \boldsymbol{x}_i' = \sum_{i=1}^{N} \boldsymbol{x}_i'.$$

- Thus, we obtain

$$\hat{\omega}_i = \left(\sum_{i=1}^{N} \boldsymbol{x}_i' \right) \left(\sum_{i \in B} \boldsymbol{x}_i \boldsymbol{x}_i' / v_i \right)^{-1} \boldsymbol{x}_i / v_i. \tag{5}$$

- If $\sum_{i=1}^{N} \boldsymbol{x}_i$ is unknown, we can estimate it from sample A to get

$$\hat{\omega}_i = \left(\sum_{i \in A} w_i^{(A)} \boldsymbol{x}_i' \right) \left(\sum_{i \in B} \boldsymbol{x}_i \boldsymbol{x}_i' / v_i \right)^{-1} \boldsymbol{x}_i / v_i . \qquad (6)$$

- The final PS weights $\hat{\omega}_i$ in (6) satisfies

$$\sum_{i \in B} \hat{\omega}_i y_i = \sum_{i \in A} w_i^{(A)} \hat{y}_i , \qquad (7)$$

where $\hat{y}_i = \boldsymbol{x}_i' \hat{\boldsymbol{\beta}}$ and

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i \in B} \boldsymbol{x}_i \boldsymbol{x}_i' / v_i \right)^{-1} \sum_{i \in B} \boldsymbol{x}_i y_i / v_i .$$

- Equation (7) gives a dual relationship between the regression weighting and the regression mass imputation. This condition is called the self-efficiency condition (Wang and Kim, 2021).

Model calibration (Wu and Sitter, 2001)

- If the outcome model is

$$Y_i = m(\boldsymbol{x}_i; \beta) + e_i,$$

with $e_i \sim (0, \sigma^2)$, then we can use the working model directly for PS estimation: Minimize

$$Q(\omega) = \sum_{i \in B} {\omega_i}^2$$

subject to

$$\sum_{i \in B} \omega_i \left[1, m(\boldsymbol{x}_i; \hat{\beta})\right] = \sum_{i \in A} w_i^{(A)} \left[1, m(\boldsymbol{x}_i; \hat{\beta})\right]. \tag{8}$$

- The uncertainty of $\hat{\beta}$ is asymptotically negligible.

Non-parametric calibration

- If

$$E(Y \mid \boldsymbol{x}) \in \mathrm{span}\{1, b_1(\boldsymbol{x}), \cdots, b_L(\boldsymbol{x})\},$$

  then, instead of (8), we may use

$$\sum_{i \in B} \omega_i \left[1, b_1(\boldsymbol{x}_i), \cdots, b_L(\boldsymbol{x}_i)\right] = \sum_{i \in A} w_i^{(A)} \left[1, b_1(\boldsymbol{x}_i), \cdots, b_L(\boldsymbol{x}_i)\right] \quad (9)$$

  in the calibration estimation. Calibration using the basis functions is called indirect model calibration.

- No need to estimate $\boldsymbol{\beta}$ in $E(Y \mid \boldsymbol{x}; \boldsymbol{\beta}) = \sum_{k=0}^{L} \beta_k b_k(\boldsymbol{x})$.
- The dimension L may increase with the sample size. In this case, some regularization method can be used to choose L.
- For example, Montanari and Ranalli (2005) used Neural Network model and Breidt et al. (2005) used penalized Spline model for nonparametric calibration estimation.

Example: PS weighting using Kernel ridge regression (Wang and Kim, 2023)

- Step 1: Use the kernel ridge regression method

$$\min_{g \in \mathcal{H}_k} \sum_{i \in B} \{y_i - m(\boldsymbol{x}_i)\}^2 + \lambda \|m\|_{\mathcal{H}_k}$$

  to get $\hat{m}(\boldsymbol{x})$ as a nonparametric estimator for $E(Y \mid \boldsymbol{x})$, where $\mathcal{H}_k$ is a reproducing kernel Hilbert space generated by kernel k.

- Step 2: Find the final PS weights to satisfy the self-efficiency:

$$\sum_{i \in B} \hat{\omega}_i y_i = \sum_{i \in A} w_i^{(A)} \hat{m}(\boldsymbol{x}_i)$$

Improvement of regression weighting

- Recall that the regression weights are

$$\hat{\omega}_i = \hat{\boldsymbol{\lambda}}' \boldsymbol{x}_i / v_i$$

where $\hat{\boldsymbol{\lambda}}$ satisfies the calibration equation in (4). That is,

$$\hat{\boldsymbol{\lambda}}' = \left(\sum_{i \in A} w_i^{(A)} \boldsymbol{x}_i \right)' \left(\sum_{i \in B} \boldsymbol{x}_i \boldsymbol{x}_i' / v_i \right)^{-1} .$$

- Note that the weight is a linear function of $\mathbf{z}_i = \boldsymbol{x}_i / v_i$. Thus, the calibration weights can take extreme values. For example, they can take negative values.

- Instead of using

$$Q_{\mathrm{LS}}(\omega) = \sum_{i \in B} {\omega_i}^2 v_i,$$

we may use other objective functions that gaurantee positive propensity weights in solving the constrained optimization problem.

- For example, we can use

$$Q_{\mathrm{EL}}(\omega) = -\sum_{i \in B} v_i \log(\omega_i)$$

or

$$Q_{\mathrm{ET}}(\omega) = \sum_{i \in B} v_i \omega_i \log(\omega_i)$$

in the optimization problem.

4. Examples

1. Information Projection approach
2. Using fractionally weighted imputation.

Information projection approach (Wang and Kim, 2021)

- Under MAR, the propensity score (PS) estimator of θ_N is defined as using

$$\hat{\theta}_{\mathrm{PS}} = N^{-1} \sum_{i \in B} \omega(\boldsymbol{x}_i) y_i,$$

where

$$\omega(\boldsymbol{x}) = \frac{1}{\mathbb{P}(\delta = 1 \mid \boldsymbol{x})}.$$

- We are interested in estimating $\omega(\boldsymbol{x})$ for efficient estimation of θ.
- Let $N = N_1 + N_0$, where $N_1 = |B|$.

Density ratio function

- Using Bayes formula, we can express

$$\begin{aligned}\mathbb{P}(\delta = 1 \mid \boldsymbol{x}) &= \frac{\pi_1 f(\boldsymbol{x} \mid I_B = 1)}{\pi_1 f(\boldsymbol{x} \mid \delta = 1) + \pi_0 f(\boldsymbol{x} \mid \delta = 0)} \\ &:= \frac{\pi_1 f_1(\boldsymbol{x})}{\pi_1 f_1(\boldsymbol{x}) + \pi_0 f_0(\boldsymbol{x})}\end{aligned}$$

  where $\pi_1 = \mathbb{P}(\delta = 1)$ and $\pi_0 = \mathbb{P}(\delta = 0)$.

- Thus, writing

$$r(\boldsymbol{x}) = \frac{f_0(\boldsymbol{x})}{f_1(\boldsymbol{x})},$$

  we can express the propensity weight as

$$\omega(\boldsymbol{x}) \equiv \frac{1}{\mathbb{P}(\delta = 1 \mid \boldsymbol{x})} = 1 + \frac{\pi_0}{\pi_1} r(\boldsymbol{x}) = 1 + \frac{N_0}{N_1} r(\boldsymbol{x}). \tag{10}$$

- Since N_0/N_1 is known, there is an one-to-one correspondence between the model for $\mathbb{P}(\delta = 1 \mid \boldsymbol{x})$ and the model for $r(\boldsymbol{x})$. The ratio $r(\boldsymbol{x})$ is called the density ratio function.

Density ratio function model

- To estimate density ratio function $r(\boldsymbol{x})$ from the sample, note that we do not even have a model for $r(\boldsymbol{x})$.
- Wang and Kim (2021) developed so-called the information projection (or I-projection) approach to find a model for $r(\boldsymbol{x})$ and then develop parameter estimation method.
- The I-projection uses the Kullback-Leibler divergence between f_0 and f_1. We wish to minimize

$$D(f_0 \parallel f_1) = \int \log\left(f_0/f_1\right) f_0 d\mu, \tag{11}$$

  w.r.t. f_0 such that $\int f_0 d\mu = 1$, and some moment constraints.
- In the spirit of indirect model calibration in (9), we can use the moment constraints in $\mathcal{H}$ where

$$\mathbb{E}(Y \mid \boldsymbol{x}) \in \text{span}\{1, b_1(\boldsymbol{x}), \cdots, b_L(\boldsymbol{x})\} := \mathcal{H}.$$

- Thus, the linear space that we are projecting on is

$$\pi_1 \int \mathbf{b}(x) f_1(x) d\mu + \pi_0 \int \mathbf{b}(x) f_0(x) d\mu = \mathbb{E}\{\mathbf{b}(X)\}, \tag{12}$$

  where $\mathbf{b}(x)$ is the basis functions in $\mathcal{H}$.

- Thus, the I-projection can be formulated as minimizing $D(f_0 \parallel f_1)$ subject to $\int f_0 d\mu = 1$ and (12).
- The solution is

$$f_0^*(x) = f_1(x) \times \frac{\exp\{\phi_1' \mathbf{b}(x)\}}{\mathbb{E}_1\left[\exp\{\phi_1' \mathbf{b}(x)\}\right]}, \tag{13}$$

  where ϕ_1 is the Lagrange multiplier satisfying (12).

- Expression (13) leads to a parametric density ratio model:

$$\log\{r^*(x)\} = \phi_0 + \phi_1 b_1(x) + \cdots + \phi_L b_L(x). \tag{14}$$

  Model (14) can be called the log-linear density ratio model.

- Once $r^*(\boldsymbol{x})$ is obtained by (14), we can apply (10) to get

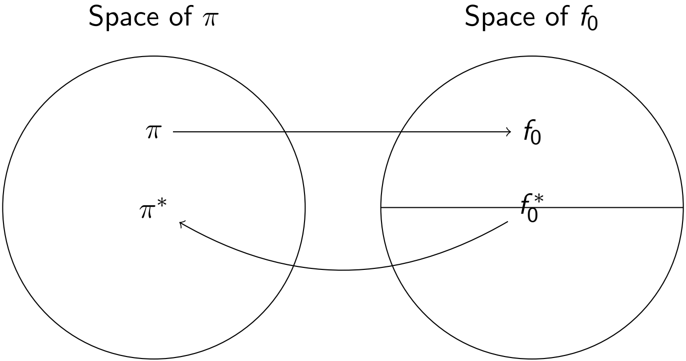
$$\omega^*(\boldsymbol{x};\phi) = 1 + \frac{N_0}{N_1}\exp\left(\phi_0 + \boldsymbol{\phi}_1'\mathbf{b}(x)\right).$$

- This is the solution to the I-projection under covariate-balancing constraints in (12).
- For parameter estimation, we use the calibration equation:

$$\sum_{i\in B}\underbrace{\left[1+\frac{N_0}{N_1}\cdot\exp\{\hat{\phi}_0+\hat{\phi}_1'\mathbf{b}(\boldsymbol{x}_i)\}\right]}_{=\hat{\omega}_i^*}[1,\mathbf{b}(\boldsymbol{x}_i)] = \left[N,\,\hat{T}_b\right], \qquad (15)$$

where

$$\hat{T}_b = \sum_{i\in A} w_i^{(A)}\mathbf{b}(\boldsymbol{x}_i).$$



The transformation from one space to another space is expressed by (10).

Remark

- For the choice of the basis functions in $\mathcal{H}$, we may use Y-variable information in sample B.
- That is, since we observe $(\boldsymbol{x}_i, y_i)$ in sample B, we can use the standard regression techniques to fit a regression model for Y from sample B:

$$y_i = \beta_0 + \sum_{k=1}^{L} \beta_k b_k(\boldsymbol{x}_i) + e_i \tag{16}$$

  with $\mathbb{E}(e_i) = 0$. Model (16) can be regarded as a "working" outcome model.
- Therefore, the basis functions obtained in (16) can be used to construct the smoothed propensity score (PS) weights using the I-projection method. The resulting PS estimator is efficient if the working model is good.

Topic 2: Using fractional weighted imputation

- Regression imputation: Use

$$\hat{y}_i = \mathbf{x}_i' \hat{\boldsymbol{\beta}}, \tag{17}$$

where

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

- Mass imputation estimator using (17):

$$\hat{\theta}_{\mathrm{RMI}} = \frac{1}{N} \sum_{i \in A} w_i^{(A)} \hat{y}_i \tag{18}$$

can be used to estimate $\theta_N = N^{-1} \sum_{i=1}^N y_i$.

- Note that we can express the regression imputation (17) as a linear function of y:
$$\hat{y}_i = \sum_{j \in B} w_{ij}^* y_j \tag{19}$$
where
$$w_{ij}^* = \mathbf{x}_i' \left(\sum_{k \in B} \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \mathbf{x}_j \tag{20}$$
and w_{ij}^* is the fractional weight assigned to donor $j \in B$ to compute $\hat{y}_i$ $(i \in A)$ in (19).
- Any imputation of the form in (19) is called the fractionally weighted imputation (FWI).

- The fractional weight w_{ij}^* in (20) satisfies

$$\begin{aligned} \sum_{j \in B} w_{ij}^* \boldsymbol{x}_j' &= \sum_{j \in B} \boldsymbol{x}_i' \left(\sum_{k \in B} \boldsymbol{x}_k \boldsymbol{x}_k' \right)^{-1} \boldsymbol{x}_j \boldsymbol{x}_j' \\ &= \boldsymbol{x}_i' \left(\sum_{k \in B} \boldsymbol{x}_k \boldsymbol{x}_k' \right)^{-1} \sum_{j \in B} \boldsymbol{x}_j \boldsymbol{x}_j' \\ &= \boldsymbol{x}_i. \end{aligned} \tag{21}$$

- Property (21) means that the fractionally weighted imputation applied to $\boldsymbol{x}_j$ in the donor set recovers $\boldsymbol{x}_i$ for each recipient $i \in A$. This property is called the reproducing property of the fractional weights.

- Note that the regression fractional weights can be used to construct the propensity score (PS) weights by writing

$$\begin{aligned}\sum_{i\in A} w_i^{(A)} \hat{y}_i &= \sum_{i\in A} w_i^{(A)} \sum_{j\in B} w_{ij}^* y_j \\ &= \sum_{j\in B} \sum_{i\in A} w_i^{(A)} w_{ij}^* y_j \\ &= \sum_{j\in B} \omega_j y_j\end{aligned}$$

where

$$\omega_j = \sum_{i\in A} w_i^{(A)} w_{ij}^* .$$

- If the fractional weights satisfy the reproducing property in (21), the PS weights ω_i satisfy the calibration property:

$$\sum_{i\in B} \omega_i \boldsymbol{x}_i = \sum_{i\in A} w_i^{(A)} \boldsymbol{x}_i .$$

Toy Example ($n_A = n_B = 3$)

Table: Data structure for data integration

	ID	Wgt	X	Y
Sample A	1	w_1	$\boldsymbol{x}_1$	?
	2	w_2	$\boldsymbol{x}_2$	?
	3	w_3	$\boldsymbol{x}_3$	?
Sample B	4	?	$\boldsymbol{x}_4$	y_4
	5	?	$\boldsymbol{x}_5$	y_5
	6	?	$\boldsymbol{x}_6$	y_6

The outcome variable is missing in sample A. In sample B, the sampling weight is missing.

Step 1: Regression fractional imputation for Sample A

Sample	ID	Donor ID	Wgt	X	X^*	Y^*
A	1	4	$w_1 w_{14}^*$	$\boldsymbol{x}_1$	$\boldsymbol{x}_4$	y_4
		5	$w_1 w_{15}^*$	$\boldsymbol{x}_1$	$\boldsymbol{x}_5$	y_5
		6	$w_1 w_{16}^*$	$\boldsymbol{x}_1$	$\boldsymbol{x}_6$	y_6
	2	4	$w_2 w_{24}^*$	$\boldsymbol{x}_2$	$\boldsymbol{x}_4$	y_4
		5	$w_2 w_{25}^*$	$\boldsymbol{x}_2$	$\boldsymbol{x}_5$	y_5
		6	$w_2 w_{26}^*$	$\boldsymbol{x}_2$	$\boldsymbol{x}_6$	y_6
	3	4	$w_3 w_{34}^*$	$\boldsymbol{x}_3$	$\boldsymbol{x}_4$	y_4
		5	$w_3 w_{35}^*$	$\boldsymbol{x}_3$	$\boldsymbol{x}_5$	y_5
		6	$w_3 w_{36}^*$	$\boldsymbol{x}_3$	$\boldsymbol{x}_6$	y_6

Step 2: Regression PS weights for Sample B

Sample	ID	Donor ID	Wgt	X	X^*	Y^*
A	1	4	$w_1 w_{14}^*$	$\boldsymbol{x}_1$	$\boldsymbol{x}_4$	y_4
		5	$w_1 w_{15}^*$	$\boldsymbol{x}_1$	$\boldsymbol{x}_5$	y_5
		6	$w_1 w_{16}^*$	$\boldsymbol{x}_1$	$\boldsymbol{x}_6$	y_6
	2	4	$w_2 w_{24}^*$	$\boldsymbol{x}_2$	$\boldsymbol{x}_4$	y_4
		5	$w_2 w_{25}^*$	$\boldsymbol{x}_2$	$\boldsymbol{x}_5$	y_5
		6	$w_2 w_{26}^*$	$\boldsymbol{x}_2$	$\boldsymbol{x}_6$	y_6
	3	4	$w_3 w_{34}^*$	$\boldsymbol{x}_3$	$\boldsymbol{x}_4$	y_4
		5	$w_3 w_{35}^*$	$\boldsymbol{x}_3$	$\boldsymbol{x}_5$	y_5
		6	$w_3 w_{36}^*$	$\boldsymbol{x}_3$	$\boldsymbol{x}_6$	y_6
B	4		?		$\boldsymbol{x}_4$	y_4
	5		?		$\boldsymbol{x}_5$	y_5
	6		?		$\boldsymbol{x}_6$	y_6

Step 2: Regression PS weights for Sample B

Sample	ID	Donor ID	Wgt	X	X^*	Y^*
A	1	4	$w_1 w_{14}^*$	$\boldsymbol{x}_1$	$\boldsymbol{x}_4$	y_4
		5	$w_1 w_{15}^*$	$\boldsymbol{x}_1$	$\boldsymbol{x}_5$	y_5
		6	$w_1 w_{16}^*$	$\boldsymbol{x}_1$	$\boldsymbol{x}_6$	y_6
	2	4	$w_2 w_{24}^*$	$\boldsymbol{x}_2$	$\boldsymbol{x}_4$	y_4
		5	$w_2 w_{25}^*$	$\boldsymbol{x}_2$	$\boldsymbol{x}_5$	y_5
		6	$w_2 w_{26}^*$	$\boldsymbol{x}_2$	$\boldsymbol{x}_6$	y_6
	3	4	$w_3 w_{34}^*$	$\boldsymbol{x}_3$	$\boldsymbol{x}_4$	y_4
		5	$w_3 w_{35}^*$	$\boldsymbol{x}_3$	$\boldsymbol{x}_5$	y_5
		6	$w_3 w_{36}^*$	$\boldsymbol{x}_3$	$\boldsymbol{x}_6$	y_6
B	4		?		$\boldsymbol{x}_4$	y_4
	5		?		$\boldsymbol{x}_5$	y_5
	6		?		$\boldsymbol{x}_6$	y_6

Step 2: Regression PS weights for Sample B

Sample	ID	Donor ID	Wgt	X	X^*	Y^*
A	1	4	$w_1 w_{14}^*$	$\boldsymbol{x}_1$	$\boldsymbol{x}_4$	y_4
		5	$w_1 w_{15}^*$	$\boldsymbol{x}_1$	$\boldsymbol{x}_5$	y_5
		6	$w_1 w_{16}^*$	$\boldsymbol{x}_1$	$\boldsymbol{x}_6$	y_6
	2	4	$w_2 w_{24}^*$	$\boldsymbol{x}_2$	$\boldsymbol{x}_4$	y_4
		5	$w_2 w_{25}^*$	$\boldsymbol{x}_2$	$\boldsymbol{x}_5$	y_5
		6	$w_2 w_{26}^*$	$\boldsymbol{x}_2$	$\boldsymbol{x}_6$	y_6
	3	4	$w_3 w_{34}^*$	$\boldsymbol{x}_3$	$\boldsymbol{x}_4$	y_4
		5	$w_3 w_{35}^*$	$\boldsymbol{x}_3$	$\boldsymbol{x}_5$	y_5
		6	$w_3 w_{36}^*$	$\boldsymbol{x}_3$	$\boldsymbol{x}_6$	y_6
B	4		?		$\boldsymbol{x}_4$	y_4
	5		?		$\boldsymbol{x}_5$	y_5
	6		?		$\boldsymbol{x}_6$	y_6

Step 2: Regression PS weights for Sample B

	ID	Wgt	X	Y
Sample B	4	ω_4	$\boldsymbol{x}_4$	y_4
	5	ω_5	$\boldsymbol{x}_5$	y_5
	6	ω_6	$\boldsymbol{x}_6$	y_6

$$\begin{aligned}
\omega_4 &= w_1 w_{14}^* + w_2 w_{24}^* + w_3 w_{34}^* \\
\omega_5 &= w_1 w_{15}^* + w_2 w_{25}^* + w_3 w_{35}^* \\
\omega_6 &= w_1 w_{16}^* + w_2 w_{26}^* + w_3 w_{36}^*
\end{aligned}$$

Remark

- The final PS weights ω_i are constructed to satisfy

$$\sum_{i \in B} \omega_i y_i = \sum_{i \in A} w_i^{(A)} \hat{y}_i. \tag{22}$$

  This condition is called the <u>self-efficiency</u> condition (Wang and Kim, 2021).

- If $\hat{y}_i = \sum_{j \in B} w_{ij}^* y_j$, then (22) implies

$$\omega_j = \sum_{i \in A} w_i^{(A)} w_{ij}^*$$

  and the PS weight construction takes the form of <u>weight sharing</u>.

- Therefore, the weight share method for data integration can be formulated as constructing w_{ij}^* (fractional weights) for fractionally weighted imputation in $\hat{y}_i = \sum_{j \in B} w_{ij}^* y_j$.
- The fractional weights w_{ij}^* should satisfy

$$\sum_{j \in B} w_{ij}^* = 1$$

and $w_{ij}^* \geq 0$.
- We propose a model-based fractional weighted imputation (FWI) as a new statistical tool to compute PS weights for data integration.
- Model-based FWI method was first proposed by Kim and Yang (2014) in the context of handling item nonresponse. The method can be directly applicable to data integration.

Proposal: Model-based Fractionally Weighted Imputation

- We can employ a statistical model $f(y \mid \boldsymbol{x}; \theta)$ to develop a kernel function for fractional imputation and nearest neighbor imputation.
- First, we use the training sample (sample B) to estimate θ consistently.
- Our goal is to create fractionally weighted imputation (FWI) such that we have

$$\sum_{j \in B} w_{ij}^{*} y_j \cong E(Y \mid \boldsymbol{x}_i; \hat{\theta}) = \int y f(y \mid \boldsymbol{x}_i; \hat{\theta}) dy. \tag{23}$$

- Once (23) is established, then we can use the self-efficiency in (22) to get

$$\omega_j = \sum_{i \in A} w_i^{(A)} w_{ij}^{*}$$

as the final PS weights for sample B.

Idea: importance sampling

- Importance sampling weight (kernel) for $i \in A$:

$$w_{ij}^{*} \propto \frac{f(y_j \mid \boldsymbol{x}_i; \hat{\theta})}{f(y_j \mid \delta_j = 1)} \tag{24}$$

  and $\sum_{j \in B} w_{ij}^{*} = 1$.
- To compute the denominator in (24), we use

$$f(y \mid \delta = 1) = \int f(y \mid \boldsymbol{x}; \hat{\theta}) f(\boldsymbol{x} \mid \delta = 1) d\boldsymbol{x} \cong \frac{1}{n_B} \sum_{i \in B} f(y \mid \boldsymbol{x}_i; \hat{\theta})$$

- Thus, we can use

$$w_{ij}^{*} \propto \frac{f(y_j \mid \boldsymbol{x}_i; \hat{\theta})}{\sum_{k \in B} f(y_j \mid \boldsymbol{x}_k; \hat{\theta})} \tag{25}$$

  as the fractional weights for model-based FWI.

Main Theory

- Assumption: The outcome model $f(y \mid \boldsymbol{x}; \theta)$ holds for the finite population and the sampling mechanism for sample B satisfies the ignorability and the positivity.

- Result: The fractional weights

$$w_{ij}^*(\hat{\theta}) \propto \frac{f(y_j \mid \boldsymbol{x}_i; \hat{\theta})}{\sum_{k \in B} f(y_j \mid \boldsymbol{x}_k; \hat{\theta})}$$

with $\sum_{j \in B} w_{ij}^* = 1$ satisfies

$$\sum_{j \in B} w_{ij}^*(\hat{\theta}) h(y_j) = E\{h(Y) \mid \boldsymbol{x}_i; \theta_0\} + O_p(n_B^{-1/2})$$

for any measurable $h(Y)$ and for $\hat{\theta} \xrightarrow{p} \theta_0$.

5. Doubly robust estimation

Protection from model misspecification

- Recall that the model-based PS weighted estimator can be expressed as

$$\sum_{i \in B} \hat{\omega}_i y_i \cong \sum_{i \in A} w_i^{(A)} \hat{y}_i$$

where $\hat{y}_i = \boldsymbol{x}_i' \hat{\beta}$ is justified under the outcome regression (OR) model

$$Y_i = \boldsymbol{x}_i' \boldsymbol{\beta} + e_i$$

where $e_i \sim (0, v_i)$.

- We now wish to incorporate $\hat{\pi}_i = \pi(\boldsymbol{x}_i; \hat{\phi})$ to achieve consistency under the propensity score (PS) model

$$P(\delta = 1 \mid \boldsymbol{x}) = \pi(\boldsymbol{x}; \phi).$$

Bias correction method

- Bias of regression estimator under PS model is

$$\text{Bias}\left(\sum_{i\in A} w_i^{(A)}\hat{y}_i\right) \cong \sum_{i=1}^{N} E(\hat{y}_i) - \sum_{i=1}^{N} y_i$$

- Using $\hat{\pi}_i$, we can estimate the bias from sample B:

$$\sum_{i=1}^{N} \frac{\delta_i}{\hat{\pi}_i}(\hat{y}_i - y_i)$$

- Thus, a bias-corrected regression estimator is

$$\hat{\theta}_{\text{reg,bc}} = \frac{1}{N}\sum_{i\in A} w_i^{(A)}\hat{y}_i + \frac{1}{N}\sum_{i\in B}\frac{1}{\hat{\pi}_i}(y_i - \hat{y}_i)\,. \qquad (26)$$

Doubly robustness

- Writing $\hat{\theta}_A = N^{-1}\sum_{i\in A} w_i^{(A)} y_i$, we obtain

$$
\begin{aligned}
\hat{\theta}_{\text{reg,bc}} - \hat{\theta}_A &= \frac{1}{N}\sum_{i\in A} w_i^{(A)}\left(\hat{y}_i - y_i\right) + \frac{1}{N}\sum_{i\in B}\frac{1}{\hat{\pi}_i}\left(y_i - \hat{y}_i\right) \\
&= -\frac{1}{N}\sum_{i=1}^{N}\left(w_i^{(A)} I_i^{(A)} - \hat{\pi}_i^{-1}\delta_i\right)\left(y_i - \hat{y}_i\right).
\end{aligned}
$$

- Double robustness: either the OR model or the PS model is correctly specified, then we obtain consistency.
- If $\hat{y}_i$ or $\hat{\pi}_i$ is nonparametrically estimated, then the effect of estimating nuisance parameter is no longer negligible. In this case, some additional techniques (debiased method) needs to be developed.

Dual expression

- The bias-corrected regression estimator in (26) can be written as

$$\begin{aligned}\hat{\theta}_{\mathrm{reg,bc}} &= \frac{1}{N}\sum_{i\in B}\frac{1}{\hat{\pi}_i}y_i + \frac{1}{N}\left(\hat{\mathbf{X}}_A - \hat{\mathbf{X}}_B\right)'\hat{\beta} \\ &= \frac{1}{N}\sum_{i\in B}\hat{\omega}_i y_i\end{aligned}$$

where

$$\hat{\omega}_i = \frac{1}{\hat{\pi}_i} + \left(\hat{\mathbf{X}}_A - \hat{\mathbf{X}}_B\right)'\left(\sum_{i\in B}\boldsymbol{x}_i\boldsymbol{x}_i'/v_i\right)^{-1}\boldsymbol{x}_i/v_i \qquad (27)$$

and

$$\left(\hat{\mathbf{X}}_A, \hat{\mathbf{X}}_B\right) = \left(\sum_{i\in A} w_i^{(A)}\boldsymbol{x}_i, \sum_{i\in B}\hat{\pi}_i^{-1}\boldsymbol{x}_i\right).$$

- Note that the regression weight in (27) satisfies

$$\sum_{i \in B} \hat{\omega}_i \boldsymbol{x}_i = \hat{\mathbf{X}}_A. \tag{28}$$

- In fact, the regression weight in (27) is the minimizer of

$$Q(\omega) = \sum_{i \in B} \left(\omega_i - \frac{1}{\hat{\pi}_i} \right)^2 v_i$$

subject to (28).

Internal bias calibration

- In the bias-corrected regression estimator in (26), the bias-correction term is added in an explicit form.
- We now consider an alternative approach which implement bias correction internally.
- Note that the regression estimator

$$\hat{\theta}_{\text{reg}} = \frac{1}{N} \sum_{i \in A} w_i^{(A)} \hat{y}_i$$

can take a bias correction term if

$$\sum_{i \in B} \frac{1}{\hat{\pi}_i} (y_i - \hat{y}_i) = 0. \tag{29}$$

- Thus, we have only to find a sufficient condition for (29).

- Recall that $\hat{y}_i = \boldsymbol{x}_i'\hat{\beta}$ satisfies

$$\sum_{i \in B} (y_i - \hat{y}_i)\, \boldsymbol{x}_i / v_i = \mathbf{0}.$$

- Thus, as long as $\boldsymbol{x}_i$ contains $v_i/\hat{\pi}_i$, we achieve (29), which implies that the resulting regression estimator is doubly robust.
- The final weight can be obtained by minimizing

$$Q_{\mathrm{LS}}(\omega) = \sum_{i \in B} v_i {\omega_i}^2$$

subject to (28) and

$$\sum_{i \in B} \omega_i \left(v_i/\hat{\pi}_i \right) = \sum_{i \in A} w_i^{(A)} \left(v_i/\hat{\pi}_i \right). \tag{30}$$

- Condition (30) can be called the internal bias calibration (IBC) condition.

Remark

- Instead of the squared error loss $Q_{\mathrm{LS}}(\omega)$, if we use

$$Q_{\mathrm{EL}}(\omega) = \sum_{i \in B} v_i \log(\omega_i),$$

then the IBC condition is changed to

$$\sum_{i \in B} \omega_i (v_i \hat{\pi}_i) = \sum_{i \in A} w_i^{(A)} (v_i \hat{\pi}_i). \tag{31}$$

- Multiple PS models can be considered in the IBC condition to get multiply robust PS weights.

6. Conclusion

- Propensity score (PS) weighting is a popular tool for adjusting for the selection bias in the non-probability sample.
- The PS weights can be developed either under the propensity score model or under the outcome regression model.
- When the subject-matter knowledge is available, the outcome regression model approach is more attractive.
- When constructing PS weights from an outcome regression model, we can impose a condition (such as the IBC condition) to protect against misspecification bias in the outcome regression model.

References I

Breidt, F. J., G. Claeskens, and J. D. Opsomer (2005). Model-assisted estimation for complex surveys using penalised splines. *Biometrika 92*, 831–846.

Chen, Y., P. Li, and C. Wu (2020). Doubly robust inference with nonprobability survey samples. *Journal of the American Statistical Association 115*(532), 2011–2021.

Kim, J. K. and Y. Kwon (2024). Discussion of "Exchangeability assumption in propensity-score based adjustment method for population mean estimation using non-probability samples" by Yan Li. *Survey Methodology*. Accepted for publication.

Kim, J. K. and S. Yang (2014). Fractional hot deck imputation for robust inference under item nonresponse in survey sampling. *Survey Methodology 40*(2), 211–230.

References II

Montanari, G. E. and M. G. Ranalli (2005). Nonparametric model calibration estimation in survey sampling. *J. Am. Statist. Assoc. 100*, 1429–1442.

Pfeffermann, D. and M. Sverchkov (1999). Parametric and semiparametric estimation of regression models fitted to survey data. *Sankhya B 61*, 166–186.

Wang, H. and J. K. Kim (2021). Information projection approach to propensity score estimation for handling selection bias under missing at random. Unpublished manuscript (Available at https://arxiv.org/abs/2104.13469).

Wang, H. and J. K. Kim (2023). Statistical inference using regularized m-estimation in the reproducing kernel hilbert space for handling missing data. *Annals of the Institute of Statistical Mathematics*. https://doi.org/10.1007/s10463-023-00872-8.

References III

Wu, C. and R. R. Sitter (2001). A model-calibration approach to using complete auxiliary information from survey data. *J. Am. Statist. Assoc. 96*, 185–193.