Survey Data Integration: Part 1

Jae-Kwang Kim

Iowa State University

October 9, 2023
Center for Statistical Science Peking University

Outline

1. Data integration under monotone missingness
2. Data integration under non-monotone missingness
3. Propensity score weighting method for data integration.
4. Some advanced topics

Missingness pattern: Planned missingness

Table: An example with monotone missingness structure

	X	Y
Sample A	✓	✓
Sample B	✓	

Table: An example with non-monotone missingness structure

	X	Y_1	Y_2
Sample A	✓	✓	✓
Sample B	✓	✓	
Sample C	✓		✓

Survey Integration Examples: Example 1

- US Census of housing and population
  - Short Form: 100 % sample (obtain basic demographic information)
  - Long form: about 16% sample (obtain other social and economic information as well as demographic information)
- Classical two-phase sampling problem: Calibration weighting for demographic variable to match known population counts from short form (Deming and Stephan, 1940).

Survey Integration Examples: Example 2

- Consumer Expenditure Survey (Zieschang, 1990)
  - Diary survey: Observe X, Y
  - Interview survey (quarterly): Observe X
- Two surveys are obtained independently from the same target population. (uses the same sampling frame.)
- Two estimates of X, $\hat{X}_1$ and $\hat{X}_2$, can be different because of the sampling errors.
- How to incorporate the information from the quarterly interview survey to diary survey estimate?

Survey Integration Examples: Example 3

- Canadian Survey of Employment, Payrolls and Hours (Hidiroglou, 2001)
  - A_1: Large sample drawn from a Canadian Customs and Revenue Agency administrative data file and auxiliary variables $\mathbf{x}$ observed.
  - A_2: Small sample from Statistics Canada Business Register and study variables y, number of hours worked by employees and summarized earnings, observed.

Two-phase sampling structure (for monotone missingness)

Table: A Simple Data structure for Data Integration

	X	Y
Sample A_1	✓	
Sample A_2	✓	✓

- If $A_2 \subset A_1$, then it is a classical two-phase sampling.
- If A_1 and A_2 are two independent samples, then it is sometimes called non-nested two-phase sampling.

2. Macro approach: GLS method

- Two parameters with three estimates:
  1. Survey one: Observe $\hat{X}_1$
  2. Survey two: Observe $\hat{X}_2$ and $\hat{Y}_2$
- GLS model

$$\begin{pmatrix} \hat{X}_1 \\ \hat{X}_2 \\ \hat{Y}_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \end{pmatrix} \tag{1}$$

where V is the variance-covariance matrix of $(e_1, e_2, e_3)'$

2. Macro approach: GLS method

- For classical two-phase sampling ($A_2 \subset A_1$),

$$V = \begin{pmatrix} V_{xx1} & V_{xx1} & V_{xy1} \\ V_{xx1} & V_{xx2} & V_{xy2} \\ V_{xy1} & V_{xy2} & V_{yy2} \end{pmatrix}$$

- For non-nested two-phase sampling,

$$V = \begin{pmatrix} V_{xx1} & 0 & 0 \\ 0 & V_{xx2} & V_{xy2} \\ 0 & V_{xy2} & V_{yy2} \end{pmatrix}$$

Lemma 1

Lemma

Assume that $\hat{X}_1$ and $\hat{X}_2$ are two unbiased estimators of μ_x and $\hat{Y}$ is an unbiased estimator of μ_y. Let

$$Q = \begin{pmatrix} \hat{X}_1 - \mu_x \\ \hat{X}_2 - \mu_x \\ \hat{Y} - \mu_y \end{pmatrix}' \begin{pmatrix} V(\hat{X}_1) & C(\hat{X}_1, \hat{X}_2) & C(\hat{X}_1, \hat{Y}) \\ C(\hat{X}_1, \hat{X}_2) & V(\hat{X}_2) & C(\hat{X}_2, \hat{Y}) \\ C(\hat{X}_1, \hat{Y}) & C(\hat{X}_2, \hat{Y}) & V(\hat{Y}) \end{pmatrix}^{-1} \begin{pmatrix} \hat{X}_1 - \mu_x \\ \hat{X}_2 - \mu_x \\ \hat{Y} - \mu_y \end{pmatrix}. \tag{2}$$

The optimal estimator of (μ_x, μ_y) that minimizes Q in (2) is

$$\hat{\mu}_x^* = \alpha^* \hat{X}_1 + (1 - \alpha^*) \hat{X}_2 \tag{3}$$

and

$$\hat{\mu}_y^* = \hat{Y} + B_1 \left(\hat{\mu}_x^* - \hat{X}_1 \right) + B_2 \left(\hat{\mu}_x^* - \hat{X}_2 \right) \tag{4}$$

Lemma (Cont'd)

where

$$\alpha^* = \frac{V(\hat{X}_2) - C(\hat{X}_1, \hat{X}_2)}{V(\hat{X}_1) + V(\hat{X}_2) - 2C(\hat{X}_1, \hat{X}_2)}$$

and

$$\begin{pmatrix} B_1 \\ B_2 \end{pmatrix} = \begin{pmatrix} V(\hat{X}_1) & C(\hat{X}_1, \hat{X}_2) \\ C(\hat{X}_1, \hat{X}_2) & V(\hat{X}_2) \end{pmatrix}^{-1} \begin{pmatrix} C(\hat{X}_1, \hat{Y}) \\ C(\hat{X}_2, \hat{Y}) \end{pmatrix}.$$

Proof.

Using the inverse of the partitioned matrix, we can write

$$Q = Q_1 + Q_2$$

where

$$Q_1 = \begin{pmatrix} \hat{X}_1 - \mu_x \\ \hat{X}_2 - \mu_x \end{pmatrix}' \begin{pmatrix} V(\hat{X}_1) & C(\hat{X}_1, \hat{X}_2) \\ C(\hat{X}_1, \hat{X}_2) & V(\hat{X}_2) \end{pmatrix}^{-1} \begin{pmatrix} \hat{X}_1 - \mu_x \\ \hat{X}_2 - \mu_x \end{pmatrix},$$

$$Q_2 = \left\{ \hat{Y} - E(\hat{Y} \mid \hat{X}_1, \hat{X}_2) \right\}' V_{ee}^{-1} \left\{ \hat{Y} - E(\hat{Y} \mid \hat{X}_1, \hat{X}_2) \right\},$$

$$E(\hat{Y} \mid \hat{X}_1, \hat{X}_2) = \mu_y + B_1(\hat{X}_1 - \mu_x) + B_2(\hat{X}_2 - \mu_x),$$

and $V_{ee} = V(\hat{Y}) - (B_1, B_2)\{V(\hat{X}_1, \hat{X}_2)\}^{-1}(B_1, B_2)'$.
Minimizing Q_1 with respect to μ_x gives $\hat{\mu}_x^*$ in (3) and minimizing Q_2 with respect to μ_y for given $\hat{\mu}_x^*$ gives $\hat{\mu}_y^*$ in (4). $\square$

Remark

- The optimal estimator of μ_y takes the form of the regression estimator with $\hat{\mu}_x^*$ as the control.
- Using (3), we can also express

$$\hat{\mu}_y^* = \hat{Y} - C\left(\hat{Y}, \hat{X}_2 - \hat{X}_1\right)\left\{V\left(\hat{X}_2 - \hat{X}_1\right)\right\}^{-1}(\hat{X}_2 - \hat{X}_1). \quad (5)$$

Remark

- For non-nested two-phase sampling, the optimal estimator based on $\hat{X}_2$, $\hat{Y}_2$ and $\hat{X}_1$:

$$\begin{aligned} \tilde{Y}_{opt} &= \hat{Y}_2 + B_{y \cdot x2} \left(\tilde{X}_{opt} - \hat{X}_2 \right) \\ \tilde{X}_{opt} &= \frac{V_{xx2}\hat{X}_1 + V_{xx1}\hat{X}_2}{V_{xx1} + V_{xx2}} \end{aligned}$$

  where $B_{y \cdot x2} = V_{yx2}/V_{xx2}$, $V_{xx1} = V(\hat{X}_1)$, $V_{xx2} = V(\hat{X}_2)$, $V_{yx2} = Cov(\hat{Y}_2, \hat{X}_2)$.
- Replace variances in $\tilde{Y}_{opt}$ by estimated variances to get $\hat{Y}_{opt}$ and $\hat{X}_{opt}$.

Advanced topic: Projection theory

- Let $\hat{\theta}_0$ be an unbiased estimator of θ.
- Let $\Lambda = \left\{ \hat{b}; E(\hat{b}) = 0 \right\}$ be the space of all unbiased estimators of zero.
- We consider the following class of unbiased estimators of θ:

$$\hat{\theta}_b = \hat{\theta}_0 - \hat{b} \tag{6}$$

  where $\hat{b} \in \Lambda$.
- The optimal estimator among the class in (6) is

$$\hat{\theta}_{\rm opt} = \hat{\theta}_0 - \hat{b}^*$$

  where $\hat{b}^*$ satisfies
  1. $\hat{b}^* \in \Lambda$
  2. $Cov(\hat{\theta}_0 - \hat{b}^*, \hat{b}) = 0$ for all $\hat{b} \in \Lambda$.
- The $\hat{b}^*$ satisfying the above two conditions is often called the <u>projection</u> of $\hat{\theta}_0$ onto Λ and is denoted $\hat{b}^* = \Pi(\hat{\theta}_0 \mid \Lambda)$.

Justification

- We wish to prove

$$V\left(\hat{\theta}_0 - \hat{b}\right) \geq V\left(\hat{\theta}_0 - \hat{b}^*\right). \tag{7}$$

- Note that

$$\begin{aligned} V\left(\hat{\theta}_0 - \hat{b}\right) &= V\left(\hat{\theta}_0 - \hat{b}^*\right) + V\left(\hat{b} - \hat{b}^*\right) \\ &+ 2Cov\left(\hat{\theta}_0 - \hat{b}^*, \hat{b} - \hat{b}^*\right) \end{aligned}$$

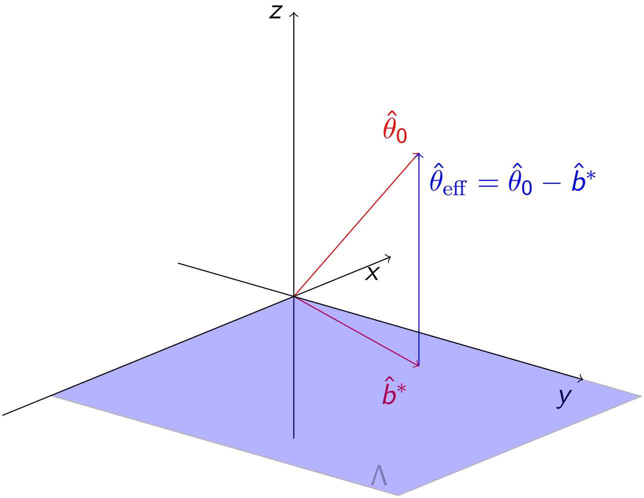
  and the covariance term is zero by the definition of $\hat{b}^*$.

- Thus, we have

$$V\left(\hat{\theta}_0 - \hat{b}\right) = V\left(\hat{\theta}_0 - \hat{b}^*\right) + V\left(\hat{b} - \hat{b}^*\right)$$

  and (7) is proved. (Pythagorean theorem)

Graphical illustration in $\mathbb{R}^3$: $\hat{b}^* = \Pi\left(\hat{\theta}_0 \mid \Lambda\right)$



3. Mass Imputation for Two-phase sampling: Basic Setup

- Two-phase sampling
  1. Phase one: observe $\mathbf{x}_i$ for $i \in A_1 \subset U$.
  2. Phase two: observe $(\mathbf{x}_i, y_i)$ for $i \in A_2 \subset A_1$.
- y_i is often expensive to measure.
- $\mathbf{x}_i$ often correlated with y_i.
- Auxiliary information of $\mathbf{x}$ in A_1 improves the estimation of $E(Y)$.

Example: simple random sampling in both phases (scalar x)

- Three estimators for two parameters:
  1. Phase one:
  $$\bar{x}_1 = \frac{1}{n_1} \sum_{i \in A_1} x_i$$
  2. Phase two:
  $$(\bar{x}_2, \bar{y}_2) = \frac{1}{n_2} \sum_{i \in A_2} (x_i, y_i) .$$
- Linear model
$$\begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \bar{y}_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \end{pmatrix} \tag{8}$$
  where $(e_1, e_2, e_3)' \sim (\mathbf{0}, \Sigma)$.
- The best estimator of $\theta = (\mu_x, \mu_y)'$ can be obtained by the (estimated) GLS method.

Example (Cont'd)

- Alternatively, if we are only interested in estimating μ_y, then we may use

$$\begin{pmatrix} \bar{x}_1 - \bar{x}_2 \\ \bar{y}_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \mu_y + \begin{pmatrix} e_1 - e_2 \\ e_3 \end{pmatrix} \tag{9}$$

- The GLS solution is then

$$\begin{aligned} \mu_y^* &= \bar{y}_2 - (\bar{x}_1 - \bar{x}_2) \frac{Cov(\bar{y}_2, \bar{x}_1 - \bar{x}_2)}{V(\bar{x}_1 - \bar{x}_2)} \\ &= \bar{y}_2 + (\bar{x}_1 - \bar{x}_2) \frac{S_{xy}}{S_{xx}} \end{aligned}$$

  where S_{xx} is the population variance of x and S_{xy} is the population covariance of x and y.

- By replacing $B = S_{xy}/S_{xx}$ by its estimate from the second-phase sample, we have

$$\bar{y}_{reg} = \bar{y}_2 + (\bar{x}_1 - \bar{x}_2) \hat{\beta}$$

  which is called the two-phase regression estimator.

Two-phase regression estimator (under SRS in both phases)

- Asymptotic Variance

$$V\left(\bar{y}_{reg,tp}\right) \doteq \left(\frac{1}{n_1} - \frac{1}{N}\right) B' S_{xx} B + \left(\frac{1}{n_2} - \frac{1}{N}\right) S_{ee}$$

- Variance comparison

$$V(\bar{y}_2) - V(\bar{y}_{reg,tp}) = \left(\frac{1}{n_2} - \frac{1}{n}\right) B' S_{xx} B \geq 0.$$

- If $Corr(x, y) \rightarrow 1$, the gain is high.

Optimal allocation

- Minimize $V\left(\bar{y}_{reg,tp}\right)$ subject to

$$C = c_0 + c_1 n_1 + c_2 n_2$$

is fixed.

- Solution:

$$\frac{n_2^*}{n_1^*} = \left(\frac{1-R^2}{R^2} \times \frac{c_1}{c_2}\right)^{1/2}$$

where $R^2 = 1 - S_e^2/S_y^2$.

Example (Cont'd)

- Two different expressions for two-phase regression estimator
  1. Calibration estimator:

$$\bar{y}_{reg} = \frac{1}{n_1} \sum_{i \in A_2} w_{ci} y_i \quad \text{where} \sum_{i \in A_2} w_{ci}(1, x_i) = (1, \bar{x}_1).$$

     Here,

$$w_{ci} = \frac{n_1}{n_2} \left\{ 1 + (\bar{x}_1 - \bar{x}_2) \frac{(x_i - \bar{x}_2)}{n_2^{-1} \sum_{i \in A_2} (x_i - \bar{x}_2)^2} \right\}.$$

  2. Imputation estimator:

$$\bar{y}_{reg} = \frac{1}{n_1} \left\{ \sum_{i \in A_2} y_i + \sum_{i \in A_1 \cap A_2^c} \hat{y}_i \right\}$$

     where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

- For estimation of mean or total, the two methods are equivalent, but they are different for estimating domain means. Which one do you prefer?

3. Mass Imputation for Two-phase sampling: Goal

1. Develop an imputation method for two-phase sampling that leads to "imputation estimator = two-phase regression estimator." Such imputation is sometimes called mass imputation.
2. Develop a replication-based variance estimation method for the above imputation procedure.

3. Mass Imputation for Two-phase sampling

- Decompose $A_1 = A_2 \cup \tilde{A}_2$ such that $A_2 \cap \tilde{A}_2 = \phi$.
  - A_2: observe $(\mathbf{x}_i, y_i)$
  - $\tilde{A}_2$: observe $\mathbf{x}_i$ only
- Wish to create y_i^*, an imputed value of y_i, for $i \in \tilde{A}_2$.
- Wish to preserve the correlation between $\mathbf{x}$ and y:

$$Corr(\mathbf{x}, y) \cong Corr(\mathbf{x}, y^*)$$

Proposed imputation method (Cont'd)

Table: Data structure for mass imputation

Sample Partition	Weight	X	Y
Phase-two sample part	w_1	x_1	y_1
	w_2	x_2	y_2
		$\vdots$	
	w_{n_2}	x_{n_2}	y_{n_2}
Remaining part	w_{n_2+1}	x_{n_2+1}	$y^*_{n_2+1}$
	w_{n_2+2}	x_{n_2+2}	$y^*_{n_2+2}$
		$\vdots$	$\vdots$
	w_{n_1}	x_{n_1}	$y^*_{n_1}$

Proposed imputation method (Cont'd)

- Let $\pi_{i2|1} = Pr(i \in A_2 \mid i \in A_1)$ be the (conditional) first-order inclusion probability for the second phase sampling. The sampling weight for the second-phase sample is $w_{i2} = w_i \pi_{i2|1}^{-1}$.
- Regression imputation: Use $y_i^* = \mathbf{x}_i' \hat{\beta}$.
- Choice of $\hat{\beta}$: Use

$$\hat{\beta} = \left(\sum_{i \in A_2} w_i \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \sum_{i \in A_2} w_i \mathbf{x}_i y_i$$

  where $\pi_{i2|1}^{-1} - 1$ is included in $\mathbf{x}_i$.
- Note that, since $\pi_{i2|1}^{-1} - 1$ is included in $\mathbf{x}_i$, we have

$$\sum_{i \in A_2} w_i (\pi_{i2|1}^{-1} - 1)(y_i - \mathbf{x}_i' \hat{\beta}) = 0. \tag{10}$$

  This is called the IBC (internal bias calibration) condition (Firth and Bennett, 1998).

IBC condition

- Two-phase regression estimator

$$\begin{aligned}\hat{Y}_{tp,reg} &= \sum_{i\in A_1} w_i \mathbf{x}_i'\hat{\beta} + \sum_{i\in A_2} w_i \frac{1}{\pi_{i2|1}}\left(y_i - \mathbf{x}_i'\hat{\beta}\right) \\ &= \text{“prediction”} + \text{“ bias correction”}\end{aligned}$$

is design consistent for any choice of $\hat{\beta}$.

- The two-phase regression estimator is model-assisted, not model-based. The regression model is used to improve the efficiency (or reduced the variance), not to obtain unbiasedness.
- If $\hat{\beta}$ satisfies (10), then we can write

$$\hat{Y}_{tp,reg} = \sum_{i\in A_2} w_i y_i + \sum_{i\in \tilde{A}_2} w_i(\mathbf{x}_i'\hat{\beta}) := \hat{Y}_{I,reg} \tag{11}$$

which is computed from mass imputation using $y_i^* = \mathbf{x}_i'\hat{\beta}$.

Domain estimation

- We may be interested in estimating domain total of y in certain domain D.
- Mass imputation provide a more efficient estimator for domain total:

$$\hat{Y}_{I,reg,D} = \sum_{i \in A_2 \cap D} w_i y_i + \sum_{i \in \tilde{A}_2 \cap D} w_i \hat{y}_i.$$

  Note that the direct estimator is

$$\hat{Y}_{tp,D} = \sum_{i \in A_2 \cap D} w_i \frac{1}{\pi_{i2|1}} y_i.$$

- Useful for small area estimation

Variance Estimation

- Replication variance estimator

$$\hat{V}_n = \sum_{k=1}^{L} c_k \left(\hat{\theta}_n^{(k)} - \hat{\theta}_n \right)^2$$

where L is the number of replication, c_k is replication factor associated with k-th replication, $\hat{\theta}_n^{(k)}$ is the k-th replicate of $\hat{\theta}_n$.

- If $\hat{\theta}_n = \sum_{i \in A} w_i y_i$, then $\hat{\theta}_n^{(k)} = \sum_{i \in A} w_i^{(k)} y_i$.
- Useful for several θ's.

Jackknife for mass imputation

- The k-th replicate of $\hat{Y}_{I,reg} = \sum_{i \in A_2} w_i y_i + \sum_{i \in \tilde{A}_2} w_i (\mathbf{x}_i' \hat{\beta})$ is

$$\hat{Y}_{I,reg}^{(k)} = \sum_{i \in A_2} w_i^{(k)} y_i + \sum_{i \in \tilde{A}_2} w_i^{(k)} (\mathbf{x}_i' \hat{\beta}^{(k)})$$

  where

$$\hat{\beta}^{(k)} = \left(\sum_{i \in A_2} w_i^{(k)} \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \sum_{i \in A_2} w_i^{(k)} \mathbf{x}_i y_i$$

- Imputed values are changed for each replication.

Application: Mass imputation for categorical data

- Let Y be a categorical with range $\{1, \cdots, K\}$.
- Assume a "working" model for $P(Y = k \mid \mathbf{x})$:

$$P(Y = k \mid \mathbf{x}) = p_k(\mathbf{x}; \beta)$$

  with $\sum_{k=1}^{K} p_k(\mathbf{x}; \beta) = 1$.
- For example, for binary y, we may use a logistic regression model

$$P(Y = 1 \mid \mathbf{x}) = \frac{\exp(\mathbf{x}'\beta)}{1 + \exp(\mathbf{x}'\beta)}.$$

- Let $\theta_k = P(Y = k)$ be the parameters of interest.
- Two-phase regression estimator of θ_k:

$$\hat{\theta}_{k,tp,reg} = \sum_{i \in A_1} w_i p_k(\mathbf{x}_i; \hat{\beta}) + \sum_{i \in A_2} w_i \pi_{i2|1}^{-1} \left\{ I(y_i = k) - p_k(\mathbf{x}_i; \hat{\beta}) \right\}.$$

  Note that $\hat{\theta}_{k,tp,reg}$ is design-consistent for θ_k, regardless of whether the working model is true or not.
- Regression Imputation estimator of θ_k:

$$\hat{\theta}_{k,I,reg} = \sum_{i \in A_2} w_i I(y_i = k) + \sum_{i \in \tilde{A}_2} w_i p_k(\mathbf{x}_i; \hat{\beta}).$$

- IBC condition:

$$\sum_{i \in A_2} w_i \left(\pi_{i2|1}^{-1} - 1 \right) \left\{ I(y_k = k) - p_k(\mathbf{x}_i; \hat{\beta}) \right\} = 0. \tag{12}$$

Remark

- For logistic regression model, we can use

$$\sum_{i \in A_2} w_i \left\{ y_i - p_k(\mathbf{x}_i; \boldsymbol{\beta}) \right\} \mathbf{x}_i = 0$$

  to estimate $\boldsymbol{\beta}$. Thus, if $\mathbf{x}_i$ includes $\pi_{i2|1}^{-1} - 1$, the IBC condition is satisfied.

- More generally, we can use an augmented regression model with

$$\sum_{i \in A_2} w_i S(\beta; \mathbf{x}_i, y_i) = 0 \tag{13}$$

  as the pseudo score equation for model parameter $\boldsymbol{\beta}$ in the working model $f(y \mid \mathbf{x}; \boldsymbol{\beta})$, where $\mathbf{x}_i$ includes $\pi_{i2|1}^{-1} - 1$ (IBC condition holds) and $S(\boldsymbol{\beta}; \mathbf{x}, y) = \partial \log f(y \mid \mathbf{x}; \boldsymbol{\beta}) / \partial \boldsymbol{\beta}$ is the score function of $\boldsymbol{\beta}$ in the parametric working model $f(y \mid \mathbf{x}; \boldsymbol{\beta})$.

Fractional mass imputation for categorical data

- First obtain $\hat{\beta}$ from (13) to satisfy IBC condition (12).
- For each unit $i \in \tilde{A}_2$, we create K imputed values

$$y_{ij}^* = j, \quad \text{with } w_{ij}^* = p_j(\mathbf{x}_i; \hat{\beta})$$

  for $j = 1, \cdots, K$.
- For variance estimation, replication method can be used:
  1. Obtain $\hat{\beta}^{(k)}$ by solving (13) with w_i replaced by $w_i^{(k)}$.
  2. The replication weights are changed to $w_{ij}^{*(k)} = p_j(\mathbf{x}_i; \hat{\beta}^{(k)})$.
- Detailed theory for mass imputation under two-phase sampling can be found in Park and Kim (2019).

4. Mass Imputation for Non-nested two-phase sampling

- Non-nested two-phase sampling
  - Survey 1: observe $\mathbf{x}_i$ for $i \in A_1$.
  - Survey 2: observe $(\mathbf{x}_i, y_i)$ for $i \in A_2$.
  - Two samples are independent.
- Wish to create mass imputation for y_i in sample A_1.
- Use a "working" regression model

$$E(y_i \mid \mathbf{x}_i) = m(\mathbf{x}_i; \beta).$$

- Mass imputation estimator (or projection estimator) of Y:

$$\hat{Y}_p = \sum_{i \in A_1} w_{i1} \hat{y}_i$$

where w_{i1} is the sampling weight for $i \in A_1$ and $\hat{y}_i = m(\mathbf{x}_i; \hat{\beta})$ where $\hat{\beta}$ is computed from A_2.

4. Mass Imputation for Non-nested two-phase sampling

- Kim and Rao (2012) show that $\hat{Y}_p$ is asymptotically design-unbiased if $\hat{\beta}$ satisfies

$$\sum_{i \in A_2} w_{i2} \left\{ y_i - m\left(\mathbf{x}_i, \hat{\beta}\right) \right\} = 0 \tag{14}$$

- Condition (14) is essentially the IBC condition for non-nested two-phase sampling.
- Note: Under condition (14),

$$\begin{aligned} \hat{Y}_p &= \sum_{i \in A_1} w_{i1} \hat{y}_i + \sum_{i \in A_2} w_{i2} \{y_i - \hat{y}_i\} \\ &= \text{“prediction”} + \text{“ bias correction”} \end{aligned}$$

- Thus, it is model-assisted, not model-based.

Theorem 1 (Kim and Rao, 2012)

- Under some regularity conditions, if $\hat{\beta}$ satisfies condition (14), we can write

$$\hat{Y}_p \cong \sum_{i \in A_1} w_{i1} m_0(x_i) + \sum_{i \in A_2} w_{i2} \{y_i - m_0(x_i)\} = \hat{P}_1 + \hat{Q}_2$$

where $m_0(x_i) = m(\mathbf{x}_i, \beta_0)$ and $\beta_0 = p\lim \hat{\beta}$ with respect to survey 2. Thus,

$$E(\hat{Y}_p) \cong \sum_{i=1}^{N} m_0(x_i) + \sum_{i=1}^{N} \{y_i - m_0(x_i)\} = \sum_{i=1}^{N} y_i.$$

and

$$V(\hat{Y}_p) \cong V(\hat{P}_1) + V(\hat{Q}_2).$$

Remark

- Model-assisted approach: Asymptotic unbiasedness of $\hat{Y}_p$ does not depend on the validity of the working model but efficiency is affected.
- Note: In the variance decomposition

$$V(\hat{Y}_p) \cong V(\hat{P}_1) + V(\hat{Q}_2) = V_1 + V_2.$$

  - V_1 is based on n_1 sample elements and V_2 is based on n_2 sample elements.
  - If $n_2 << n_1$, then $V_1 << V_2$.
  - If the working model is good, then the squared error terms $e_i^2 = \{y_i - m_0(x_i)\}^2$ are small and V_2 will also be small.

Variance estimation

- Let $e_i = y_i - \tilde{y}_i$, then the variance estimator of $\hat{Y}_p$ is

$$v_L(\hat{Y}_p) = v_1(\tilde{y}_i) + v_2(\hat{e}_i)$$

$$\begin{aligned} v_1(\tilde{z}_i) = v(\hat{Z}_1) &= \text{variance estimator for survey 1} \\ v_2(\tilde{z}_i) = v(\hat{Z}_2) &= \text{variance estimator for survey 2} \end{aligned}$$

$\hat{Z}_1 = \sum_{i \in A_1} w_{i1} z_i$, $\hat{Z}_2 = \sum_{i \in A_2} w_{i2} z_i$.

- Note $v_L(\hat{Y}_p)$ requires access to data from both surveys.
- Kim and Rao (2012) also discussed replication variance estimation for $\hat{Y}_p$.

5. Mass imputation using a non-probability training sample

- We are now interested in combining information from two samples, one with probability sampling and the other with non-probability sampling (such as voluntary sample).
- We observe X from the probability sample and observe (X, Y) from the non-probability sample. Thus, the non-probability sample is a training sample for mass imputation.

Table: Data Structure

Data	X	Y	Representativeness
A	✓		Yes
B	✓	✓	No

5. Mass imputation using a non-probability training sample

- Our parameter of interest is $\theta = E(Y)$.
- Wish to combine the two data sets to obtain an unbiased estimator of θ.
- One approach is to use mass imputation, where we use sample B as a training sample for developing the prediction model for missing Y in sample A.
- Unlike the previous case in Section 4-5, the sampling mechanism for sample B is unknown. Some additional assumptions are needed to harness the information from sample B.

Mass imputation

Mass imputation is a special case of transfer learning (in machine learning).

- How to transfer knowledge from sample B to sample A?
- Conditions for transfer learning
  1. Two samples are obtained from the same finite population.
  2. Two samples should be able to remove the selection bias (i.e. probability samples.)
  3. The measurement for two samples should be identical (i.e. use the same questionnaire.)
- If the conditions are satisfied, then a single model can be used for the two samples.

Mass Imputation using a regression model

- Regression superpopulation model

$$Y_i = m(\mathbf{x}_i; \boldsymbol{\beta}) + e_i \tag{15}$$

for some $\boldsymbol{\beta}$ with known function $m(\cdot)$, with $E(e_i \mid \mathbf{x}_i) = 0$.

- Once a consistent estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is obtained from sample B, we may use

$$\bar{y}_I = \frac{1}{N} \sum_{i \in A} w_i m(\mathbf{x}_i; \hat{\boldsymbol{\beta}}) \tag{16}$$

as the mass imputation estimator of $\theta = E(Y)$, where w_i is the sampling weight for unit $i \in A$.

How to obtain $\hat{\beta}$ in (16)?

- If $B = U$, then we can use the following estimating equation for $\boldsymbol{\beta}$

$$\sum_{i \in U} \{y_i - m(\mathbf{x}_i; \boldsymbol{\beta})\} h(\mathbf{x}_i; \boldsymbol{\beta}) = 0 \tag{17}$$

for some p-dimensional vector $h(\mathbf{x}_i; \boldsymbol{\beta})$.

- Define

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

for $i = 1, 2 \cdots, N$.

- If $\pi_i^{(B)} = P(\delta_i = 1 \mid x_i, y_i)$ is known, we may use

$$\sum_{i \in B} \frac{1}{\pi_i^{(B)}} \{y_i - m(\mathbf{x}_i, \boldsymbol{\beta})\} h(\mathbf{x}_i; \boldsymbol{\beta}) = 0 \tag{18}$$

where $\pi_i^{(B)} = P(\delta_i = 1 \mid \mathbf{x}_i, y_i)$.

- Unfortunately, we do not know $\pi_i^{(B)}$ as sample B is a non-probability sample.
- Two approaches
    1. Assume MAR: Under MAR, we may ignore $\pi_i^{(B)}$ in estimating $\boldsymbol{\beta}$.
    2. Make a model assumption for $\pi_i^{(B)}$ and use the estimated $\hat{\pi}_i^{(B)}$ in (18).
- MAR assumption (Rubin, 1976)

$$P(\delta = 1 \mid \mathbf{x}, y) = P(\delta = 1 \mid \mathbf{x}).$$

- Under MAR, we can use

$$\sum_{i \in B} \{y_i - m(\mathbf{x}_i, \boldsymbol{\beta})\} h(\mathbf{x}_i; \boldsymbol{\beta}) = 0 \tag{19}$$

to compute $\hat{\boldsymbol{\beta}}$.

Theorem 2 (Kim et al., 2021)

Assume the regression superpopulation model (15) and MAR. Under some regularity conditions, the mass imputation estimator

$$\bar{y}_I = \frac{1}{N}\sum_{i\in A} w_i m(\mathbf{x}_i;\hat{\boldsymbol{\beta}}) \tag{20}$$

satisfies

$$\bar{y}_I = \tilde{y}_I(\boldsymbol{\beta}_0) + o_p(n_B^{-1/2}) \tag{21}$$

where

$$\tilde{y}_I(\boldsymbol{\beta}) = N^{-1}\sum_{i\in A} w_i m(\mathbf{x}_i;\boldsymbol{\beta}) + n_B^{-1}\sum_{i\in B}\{y_i - m(\mathbf{x}_i;\boldsymbol{\beta})\}\, h(\mathbf{x}_i;\boldsymbol{\beta})'\mathbf{c}^*,$$

$$\mathbf{c}^* = \left[n_B^{-1}\sum_{i\in B}\dot{m}(\mathbf{x}_i;\boldsymbol{\beta}_0)h'(\mathbf{x}_i;\boldsymbol{\beta}_0)\right]^{-1} N^{-1}\sum_{i=1}^{N}\dot{m}(\mathbf{x}_i;\boldsymbol{\beta}_0),$$

$\boldsymbol{\beta}_0$ is the true value of $\boldsymbol{\beta}$ in (15), and $\dot{m}(\mathbf{x};\boldsymbol{\beta}) = \partial m(\mathbf{x};\boldsymbol{\beta})/\partial\boldsymbol{\beta}$.

Also,

$$E\{\tilde{y}_I(\boldsymbol{\beta}_0) - \bar{y}_N\} = 0, \tag{22}$$

and

$$\begin{aligned} V\left\{\tilde{y}_I(\boldsymbol{\beta}_0) - \bar{y}_N\right\} &= V\left\{N^{-1}\sum_{i\in A} w_i m(\mathbf{x}_i;\boldsymbol{\beta}_0) - N^{-1}\sum_{i\in U} m(\mathbf{x}_i;\boldsymbol{\beta}_0)\right\} \\ &+ E\left[n_B^{-2}\sum_{i\in B} E\left(e_i^2 \mid \mathbf{x}_i\right)\left\{h(\mathbf{x}_i;\boldsymbol{\beta}_0)'\mathbf{c}^*\right\}^2\right], \end{aligned} \tag{23}$$

where $e_i = y_i - m(\mathbf{x}_i;\boldsymbol{\beta}_0)$.

Example

- Under the special case of linear model

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

with $e_i \sim (0, \sigma_e^2)$, we can use $\hat{y}_i = \mathbf{x}_i' \hat{\boldsymbol{\beta}}$ with $\hat{\boldsymbol{\beta}} = \left(\sum_{i \in B} \mathbf{x}_i \mathbf{x}_i'\right)^{-1} \sum_{i \in B} \mathbf{x}_i y_i$ to construct regression mass imputation.

- If we assume SRS for sample A, we obtain

$$V\left(\bar{y}_{I,reg}\right) = V\left(\frac{1}{n_A}\sum_{i \in A} \mathbf{x}_i \boldsymbol{\beta}\right) + V\left(\frac{1}{n_B}\sum_{i \in B} e_i \mathbf{x}_i' \mathbf{c}^*\right) \tag{24}$$

where

$$\mathbf{c}^* = \left(\frac{1}{n_B}\sum_{i \in B} \mathbf{x}_i \mathbf{x}_i'\right)^{-1} \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i .$$

Example (Cont'd)

- If $\mathbf{x}_i' = (1, x_i)$, then the asymptotic variance in (24) reduces to

$$V\left(\hat{\theta}_{I,reg}\right) = \frac{1}{n_A}\beta_1^2\sigma_x^2 + \frac{1}{n_B}\sigma_e^2 + \frac{(\bar{x}_N - \bar{x}_B)^2}{\sum_{i\in B}(x_i - \bar{x}_B)^2}\sigma_e^2.$$

- If sample B were an independent random sample of size n_B, then the third term would of order $O(n_B^{-2})$ and is negligible. However, as sample B is a non-probability sample, the third term is not negligible.

Variance estimation

- For variance estimation of the mass imputation estimator (20), we have only to estimate the variance of the linearized estimator $\tilde{y}_I(\boldsymbol{\beta}_0)$ in (21). Since the variance formula can be written as

$$V\left\{\tilde{y}_I(\boldsymbol{\beta}_0) - \bar{y}_N\right\} = V_A + V_B$$

where

$$\begin{aligned} V_A &= V\left\{N^{-1}\sum_{i\in A} w_i m(\mathbf{x}_i;\boldsymbol{\beta}_0) - N^{-1}\sum_{i\in U} m(\mathbf{x}_i;\boldsymbol{\beta}_0)\right\} \\ V_B &= E\left[n_B^{-2}\sum_{i\in B} E\left(e_i^2 \mid x_i\right)\left\{h(\mathbf{x}_i;\boldsymbol{\beta}_0)'^{*}\right\}^2\right], \end{aligned}$$

we can estimate V_A and V_B separately.

- To estimate $\hat{V}_A$, we can use

$$\hat{V}_A = N^{-2} \sum_{i \in A} \sum_{j \in A} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} w_i m(\mathbf{x}_i; \hat{\beta}) w_j m(\mathbf{x}_j; \hat{\beta}).$$

  where π_{ij} is the joint inclusion probability for unit i and j, which is assumed to be positive.

- To estimate V_B, we can use

$$\hat{V}_B = n_B^{-2} \sum_{i \in B} \hat{e}_i^2 \left\{ h(\mathbf{x}_i; \hat{\beta})' \hat{\mathbf{c}}^* \right\}^2, \quad (25)$$

  where $\hat{e}_i = y_i - m(\mathbf{x}_i; \hat{\beta})$ and

$$\hat{\mathbf{c}}^* = \left[n_B^{-1} \sum_{i \in B} \dot{m}(\mathbf{x}_i; \hat{\beta}) h'(\mathbf{x}_i; \hat{\beta}) \right]^{-1} N^{-1} \sum_{i \in A} w_i \dot{m}(\mathbf{x}_i; \hat{\beta})$$

- Hence, the variance of $\bar{y}_{I,reg}$ can be estimated by $\hat{V}(\bar{y}_{I,reg}) = \hat{V}_A + \hat{V}_B$.

6. An illustrative example: NRI survey

- National Resource Inventory (NRI): Large-scale cross-sectional and longitudinal survey of land use and natural resources, sponsored by NRCS (Natural Resources Conservation Service) at USDA.
- Two-phase sampling
  1. Phase one: 1997 NRI
  2. Phase two: annual NRI
- Multi-mode data collection
  - Photo-interpretation
  - Auxiliary materials
  - Local NRCS
- Multi-purpose survey

Sampling in time

1997 Foundation sample	2000	2001	2002	2003	2004	2005	2006	2007
X	X	X	X	X	X	X	X	X
X		X						
X			X					
X				X				
X					X		X	
X						X		X
X								
X								
X								
X								

Estimation Inputs

- Survey data
  - Segment data
  - Point data
  - Selection probabilities
- Geographic control data
  - GIS surface area
  - GIS large streams
  - GIS acres of large water
  - GIS acres of federal
- Administrative control data
  - CRP acres

Goals

- Easy-to-tabulate final data set: contains all information
- "Estimate" agree with "known" controls.
- "Estimate" for 1997 values equal to the estimates from the foundation (1997) sample.
- "Best" estimates for key variables in the state level
- Reasonable estimates for small areas
- Relatively simple
- Variance estimates

Procedures

- Data preparation for 2003-style estimation
  - Turn segment geometry into acres
  - Impute variables that were collected in the past and are no longer collected
  - Reconcile multiple report for same variable in same year
- Estimation of control totals via GLS
- Imputation
  - Interpolate and extrapolate planned missing data
  - Pseudo point imputation to represent segment and control total information at the point level
  - Retention points
- Weighting adjustment
  - Change point and no-change point
  - Ratio and raking adjustment
- Variance estimation using replicates

General linear square (GLS) estimation

- For example, consider

	00	01	02	03
P0(core)	X	X	X	X
P1		X		
P2			X	
P3				X
	θ_{00}	θ_{01}	θ_{02}	θ_{03}

  where θ_t=population total at time t
- 7 estimators for 4 parameters:

$$\hat{Y} = (\hat{Y}_{0,00}, \hat{Y}_{0,01}, \hat{Y}_{1,01}, \hat{Y}_{0,02}, \hat{Y}_{2,02}, \hat{Y}_{0,03}, \hat{Y}_{3,03})^T$$

  where $\hat{Y}_{p,t}$=panel total estimate from panel p at time t $E(\hat{Y}_{p,t}) = \theta_t$
- GLS method can be used to combine the information.

Estimates

- GLS estimate for θ:
$$\hat{\theta} = (X^T V^{-1} X)^{-1} X^T V^{-1} \hat{Y}$$
- Thus, the GLS estimator combines the information.
- Different choice of V can make different GLS estimator.

Imputation

- Imputation: Fill in missing values by a (set of) plausible value(s).
- Missing structure in NRI
  1. Planned missingness in rotation sampling scheme
  2. Missingness due to multi-mode survey (segment data vs point data)
  3. Missingness due to two-phase sampling (Mass imputation)

Imputation for planned missingness

1997	2000	2001	2002	2003
X	X	X	X	X
X	I	X	I	E
X	I	I	X	E
X	I	I	I	X
X				
X				
X				
X				

X: observed, E: extrapolation, I: interpolation

Example (interpolation)

	1997 (j=1)	2000 (j=2)	2001 (j=3)	2002 (j=4)	2003 (j=5)
2003 panel Large urban	80	(80)	(80)	(85)	85
2003 panel Small water	14	(14)	(14)	(10)	10
GLS estimates Large urban	140	150	160	180	200

- Minimize the number of changes
- Use GLS estimates

Imputation for incorporating the segment level information

- Segment data

	Urban	Roads	Small Water	Total Acres
1982	32	2.9	10	160
1987	32	2.9	10	160
1992	32	2.9	10	160
1997	50	2.7	10	160

- Point data

	Point 1	Point 2	Point 3
1982	Pasture	Corn	Soybeans
1987	Pasture	Soybeans	Corn
1992	Pasture	Corn	Soybeans
1997	Pasture	Soybeans	Corn

Imputation for incorporating the segment level information

- Imputed points (pseudo points)

	No. 1	No. 2	No 3.	No.4	No. 5
1982	S. Water	Urban	Soybeans	Roads	Roads
1987	S. Water	Urban	Corn	Roads	Roads
1992	S. Water	Urban	Soybeans	Roads	Roads
1997	S. Water	Urban	Urban	Roads	Urban
Acres	10	32	17.8	2.7	0.2

- Real Points

	Point 1	Point 2	Point 3
1982	Pasture	Corn	Soybeans
1987	Pasture	Soybeans	Corn
1992	Pasture	Corn	Soybeans
1997	Pasture	Soybeans	Corn
Acreas	32.433	32.433	32.433

Imputation for two-phase sampling (mass imputation)

1997	2000	2001	2002	2003
X	X	X	X	X
X	I	X	I	E
X	I	I	X	E
X	I	I	I	X
X	M	M	M	M
X	M	M	M	M
X	M	M	M	M
X	M	M	M	M

X: observed, E: extrapolation, I: interpolation, M: mass imputation

Imputation Schemes - Summary

Sample	year (p)	year (T)
Core	Control coveruse	
	Noncontrol Coveruse	Change Points: Weighting adjustment
		No change Point: donors for imputation
Supplement sample 1	Control coveruse	Pseudo point imputation
	Noncontrol coveruse	Donor imputation
Supplement sample 2	Interpolation	Observed
Remaining sample	Mass imputation	

Weighting adjustment

- Nonresponse weighting adjustment: Change point
- Calibration:
  - adjust the original weights to satisfy the benchmarking constraint
  - raking-ratio estimation, regression weighting estimation

Variance estimation

- Replication variance estimation
  - Delete-a-group jackknife
  - Incorporate the variance due to two-phase sampling and weighting.

7. Conclusion

- Two-phase sampling is a cost-effective method of estimation for samples with missingness by design.
- Mass imputation can be developed to implement the two-phase regression estimation. Significant efficiency gains are achieved for domain estimation.
- In two-phase sampling, both samples are obtained from probability sampling designs. So, design consistency can be obtained without relying on the model assumption. (Model-assisted)
- Using a non-probability sample data set as a training set for prediction, we can implement mass imputation for survey sample data under some strong model assumptions. (Model-based)

References I

Deming, W. E. and F. F. Stephan (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics 11*, 427–444.

Firth, D. and K. Bennett (1998). Robust models in probability sampling. *J. R. Statist. Soc. B 60*, 3–21.

Hidiroglou, M. (2001). Double sampling. *Survey Methodol. 27*, 143–54.

Kim, J. K., S. Park, Y. Chen, and C. Wu (2021). Combining non-probability and probability survey samples through mass imputation. *J. R. Statist. Soc. A 184*, 941–963.

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

Park, S. and J. K. Kim (2019). Mass imputation for two-phase sampling. *Journal of the Korean Statistical Society 48*, 578–592.

References II

Rubin, D. B. (1976). Inference and missing data. *Biometrika 63*, 581–90.

Zieschang, K. (1990). Sample weighting method and estimation of totals in the consumer expenditure survey. *J. Am. Statist. Assoc. 85*, 986–1001.