Survey Data Integration: Part 2

Jae-Kwang Kim

Iowa State University

October 9, 2023

Center for Statistical Science Peking University

Outline

1. Introduction
2. Prediction under non-monotone missing patterns
3. Statistical matching
4. Measurement error model approach 1: Correcting measurement bias using internal calibration sample
5. Measurement error model approach 2: Correcting measurement bias using external calibration sample
6. Conclusion

1. Introduction

- We are interested in combining information from several probability samples.
- Under non-monotone missingness, combining information effectively can be challenging.
- The measurement error model approach can be used to combine two independent surveys under heterogeneity.

Example 1: Sampling in time

Sample	$t = 1$	$t = 2$	
A	O	O	$\rightarrow$ core panel part (detecting change)
B	O		supplemental panel survey (cross sectional)
C		O	

Sample A : $\bar{y}_{1A}$, $\bar{y}_{2A}$,
Sample B : $\bar{y}_{1B}$,
Sample C : $\bar{y}_{2C}$

Two Time Periods GLS

$$\begin{pmatrix} \bar{y}_{1B} \\ \bar{y}_{1A} \\ \bar{y}_{2A} \\ \bar{y}_{2C} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \bar{y}_{1,N} \\ \bar{y}_{2,N} \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{pmatrix}$$

$$V \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{pmatrix} = \begin{bmatrix} n_B^{-1} & 0 & 0 & 0 \\ 0 & n_A^{-1} & n_A^{-1}\rho & 0 \\ 0 & n_A^{-1}\rho & n_A^{-1} & 0 \\ 0 & 0 & 0 & n_C^{-1} \end{bmatrix} \sigma^2$$

Composite estimator $\hat{\theta} = (\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{y}$

1. Introduction

Example 2 Split questionnaire design

- Split the original sample into three groups
- In group 1, ask (x, y_1, y_2)
- In group 2, ask (x, y_1)
- In group 3, ask (x, y_2)
- Often used to reduce the response burden (and improve the quality of the survey responses).

Table: Data structure

	X	Y_1	Y_2
Group 1	✓	✓	✓
Group 2	✓	✓	
Group 3	✓		✓

Example 3 Mixed mode survey

Table: Data structure

	X	Y_1	Y_2
Sample A	✓	✓	
Sample B	✓		✓

- Y_1: measurement of Y under mode A.
- Y_2: measurement of Y under mode B.
- Y_1 and Y_2 are never jointly observed. That is, Y_2 is a counterfactual outcome of Y_1.
- It is a measurement error model problem.

1. Introduction

Two types of data structure

- Type 1: Full joint modeling is possible without additional identifying assumptions (Example 1, 2)
- Type 2: Some extra assumptions are needed to make a joint model (Example 3).

1. Introduction

Suppose that (X, Y_1, Y_2) follows a multivariate normal distribution

$$\begin{pmatrix} X \\ Y_1 \\ Y_2 \end{pmatrix} \sim N \left[\begin{pmatrix} \mu_x \\ \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_{xx} & \sigma_{x1} & \sigma_{x2} \\ & \sigma_{11} & \sigma_{12} \\ & & \sigma_{22} \end{pmatrix} \right]$$

- Under the setup of Example 2, all parameters can be estimated from the data (type 1 structure).
- Under the setup of Example 3, σ_{12} cannot be estimated directly.

2. Prediction under non-monotone missing patterns

Basic Steps

1. Model specification
2. Parameter estimation
3. Prediction (i.e. mass imputation)
4. Uncertainty quantification

2.1 Model specification

- The model is a description of the population structures (or relationships) among the survey items.
- The goal of the modeling is to predict the unobserved part of the data using the observed part. The prediction model is $f(Y_{mis} \mid Y_{obs})$, where (Y_{obs}, Y_{mis}) is the (observed, missing) part of the data.
- Once a model is specified, we only have to estimate the parameters of the model.
- Some prior information can be used to build a good model.

Example 4

- Suppose that we have the following data sources
  1. Sample A: observe (Y_1, Y_2, Y_3)
  2. Sample B: observe (Y_1, Y_2)
- Our goal is to predict Y_3 for sample B. Thus, we need to build a model for $f(Y_3 \mid Y_1, Y_2)$.
- Note that we do not need to specify the marginal distribution of (Y_1, Y_2). All we need is the model for the conditional distribution $f(Y_{mis} \mid Y_{obs})$.
- If, in addition to sample A and B, there is another sample, sample C, such that we observe (Y_2, Y_3), then we also need to specify $f(Y_1 \mid Y_2, Y_3)$.
- In this case, the two models, $f(Y_3 \mid Y_1, Y_2)$ and $f(Y_1 \mid Y_2, Y_3)$, may not be compatible with each other.

2.1 Model specification

- Two conditional models, $f(Y_1 \mid Y_2)$ and $f(Y_2 \mid Y_1)$, are called compatible if there exists a joint model such that the conditional models are obtained from the joint model.

$$f(Y_1 \mid Y_2) = \frac{f(Y_1, Y_2)}{\int f(Y_1, Y_2) dY_1} \quad f(Y_2 \mid Y_1) = \frac{f(Y_1, Y_2)}{\int f(Y_1, Y_2) dY_2}.$$

- Thus, in the previous example, we should specify $f(Y_1, Y_3 \mid Y_2)$ first and derive each conditional distribution from the joint distribution.
- We may use

$$f(Y_1, Y_3 \mid Y_2) = f(Y_1 \mid Y_2) f(Y_3 \mid Y_1, Y_2)$$

to specify the joint distribution.

Back to Example 3 (Mixed mode surveys)

- Two models (Y_1: gold standard)
  - $f(Y_1 \mid X)$: process model, structure model
  - $f(Y_2 \mid Y_1, X)$: data model for measurement Y_2, measurement model
- For example, we may consider

$$Y_1 = \beta_0 + \beta_1 X + e$$

for $f(Y_1 \mid x)$ and consider

$$Y_2 = \alpha_0 + \alpha_1 Y_1 + \alpha_2 X + u$$

for $f(Y_2 \mid Y_1)$.

- Combining two models, we obtain

$$Y_2 = \alpha_0 + \alpha_1\beta_0 + (\alpha_1\beta_1 + \alpha_2)X + \alpha_1 e + u$$

We can estimate (β_0, β_1) from sample A, but we cannot identify $(\alpha_0, \alpha_1, \alpha_2)$ from sample B.

- The model is not identifiable under the data structure in Example 3.
- If we assume $\alpha_2 = 0$, then the model is identified and we can estimate (α_0, α_1) from sample B.
- That is, to identify the model, we make an assumption that the measurement errors are invariant with respect to other covariates:

$$f(Y_2 \mid Y_1, X) = f(Y_2 \mid Y_1) \tag{1}$$

Assumption (1) is often called the non-differentiable measurement error assumption.

2.2 Parameter estimation

- Once a model is specified, we need to estimate the parameters from the data.
- Two approaches
  1. GLS-type approach
  2. (Modified) EM algorithm

Example 5

- Data Structure

Table: Data structure

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

- Model Specification

$$f(Y_1, Y_2 \mid X) = f(Y_1 \mid X; \theta_1) f(Y_2 \mid X, Y_1; \theta_2)$$

for some unknown parameter θ_1 and θ_2.

- From sample A, we can obtain $\hat{\theta}_{1,A}$ and $\hat{\theta}_{2,A}$
- From sample B, we can obtain $\hat{\theta}_{1,B}$.
- How to combine these?

GLS-type approach

- GLS model

$$\begin{pmatrix} \hat{\theta}_{1,A} \\ \hat{\theta}_{2,A} \\ \hat{\theta}_{1,B} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \end{pmatrix} \quad (2)$$

where $V\{(e_1, e_2, e_3)'\} = \mathrm{diag}\{V(\hat{\theta}_{1,A}), V(\hat{\theta}_{2,A}), V(\hat{\theta}_{1,B})\}$.

- Best linear unbiased estimator

$$\hat{\theta}_1^* = \boldsymbol{\alpha}\hat{\theta}_{1,A} + (I - \boldsymbol{\alpha})\,\hat{\theta}_{1,B}$$

where $\boldsymbol{\alpha} = \left\{V(\hat{\theta}_{1,A}) + V(\hat{\theta}_{1,B})\right\}^{-1} V(\hat{\theta}_{1,B})$.

Remark

- If the data structure is non-monotone missing structure, then the above GLS method is less straightforward.
- For example, suppose that we have the following structure.

Table: Data structure

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

- In this case, the same model can be specified:

$$f(Y_1, Y_2 \mid X) = f_1(Y_1 \mid X; \theta_1) f_2(Y_2 \mid X, Y_1; \theta_2)$$

- However, it is not easy to estimate these parameters from sample C.

(Modified) EM algorithm

1. Standardize the sampling weights for each sample such that $\sum_{i \in A} w_{i,A} = n_A$, $\sum_{i \in B} w_{i,B} = n_B$, and $\sum_{i \in C} w_{i,C} = n_C$.
2. Apply EM algorithm for the weighted sample. Let $S = A \cup B \cup C$.
    - E-step: Compute the conditional expectation of the pseudo log-likelihood:

$$
\begin{aligned}
Q_1(\theta_1 \mid \theta^{(t)}) &= \sum_{i \in S} w_i E\left\{\log f_1(y_{1i} \mid x_i; \theta_1) \mid x_i, y_{i,obs}; \theta^{(t)}\right\} \\
Q_2(\theta_2 \mid \theta^{(t)}) &= \sum_{i \in S} w_i E\left\{\log f_2(y_{2i} \mid x_i, y_{1i}; \theta_2) \mid x_i, y_{i,obs}; \theta^{(t)}\right\}
\end{aligned}
$$

      where $y_{i,obs}$ is the observed part of (y_{1i}, y_{2i}).
    - M-step: Update the parameters by finding the maximizer of $Q_1(\theta_1 \mid \theta^{(t)})$ and $Q_2(\theta_2 \mid \theta^{(t)})$ with respect to θ_1 and θ_2.

- Note that

$$
\begin{aligned}
Q_1(\theta_1 \mid \theta^{(t)}) &= \sum_{i \in S} w_i E\left\{\log f_1(y_{1i} \mid x_i; \theta_1) \mid x_i, y_{i,obs}; \theta^{(t)}\right\} \\
&= \sum_{i \in A} w_{i,A} \log f_1(y_{1i} \mid x_i; \theta_1) + \sum_{i \in B} w_{i,B} \log f_1(y_{1i} \mid x_i; \theta_1) \\
&+ \sum_{i \in C} w_{i,C} E\left\{\log f_1(Y_{1i} \mid x_i; \theta_1) \mid x_i, y_{i2}; \theta^{(t)}\right\}
\end{aligned}
$$

where the conditional expectation in sample C is with respect to

$$
f(Y_1 \mid X, Y_2; \theta^{(t)}) = \frac{f_1(Y_1 \mid X; \theta_1^{(t)}) f_2(Y_2 \mid X, Y_1; \theta_2^{(t)})}{\int f_1(Y_1 \mid X; \theta_1^{(t)}) f_2(Y_2 \mid X, Y_1; \theta_2^{(t)}) dY_1}.
$$

- Similarly, we have

$$
\begin{aligned}
Q_2(\theta_2 \mid \theta^{(t)}) &= \sum_{i \in S} w_i E \left\{ \log f_2(y_{2i} \mid x_i, y_{1i}; \theta_2) \mid x_i, y_{i,obs}; \theta^{(t)} \right\} \\
&= \sum_{i \in A} w_{i,A} \log f_2(y_{2i} \mid x_i, y_{1i}; \theta_2) \\
&+ \sum_{i \in B} w_{i,B} E \left\{ \log f_2(Y_{2i} \mid x_i, y_{1i}; \theta_2) \mid x_i, y_{i1}; \theta^{(t)} \right\} \\
&+ \sum_{i \in C} w_{i,C} E \left\{ \log f_2(y_{2i} \mid x_i, Y_{1i}; \theta_2) \mid x_i, y_{i2}; \theta^{(t)} \right\},
\end{aligned}
$$

where the conditional expectation in sample B is with respect to $f_2(Y_2 \mid X, Y_1; \theta_2^{(t)})$.

2.3 Prediction (= Mass Imputation)

- Once the parameters for the specified model are estimated, then we can predict unobserved items in the data.
- For the data structure in Example 5,

Table: Data structure

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

we impute Y_2 for sample B and impute Y_1 for sample C.

- The imputation model for Y_2 in sample B is $f_2(Y_2 \mid X, Y_1; \hat{\theta}_2)$. Also, the imputation model for Y_1 in sample C is

$$f(Y_1 \mid X, Y_2; \hat{\theta}) = \frac{f_1(Y_1 \mid X; \hat{\theta}_1) f_2(Y_2 \mid X, Y_1; \hat{\theta}_2)}{\int f_1(Y_1 \mid X; \hat{\theta}_1) f_2(Y_2 \mid X, Y_1; \hat{\theta}_2) dY_1}. \quad (3)$$

3. Statistical Matching

Statistical Matching: Combining two surveys

- Survey Items
  - X: demographic variables
  - Y_1: Health variables
  - Y_2: Social economic variables
- Two different surveys
  - Survey A: Health-related survey (Observe X and Y_1)
  - Survey B: Socio-Economic survey (Observe X and Y_2)
- Interested in fitting a regression of Y_1 (e.g. Obesity) on X and Y_2 using two surveys.
- Two samples should be obtained from the same finite population.

Classical Approach

- We want to create Y_1 for each element in sample B by finding a "statistical twin" from sample A.
- Often based on the assumption that Y_1 and Y_2 are conditionally independent, conditional on X. That is,

$$Y_1 \perp Y_2 \mid X$$

- Under CI (Conditional Independence) assumption, we have

$$f(y_1 \mid x, y_2) = f(y_1 \mid x)$$

and the "statistical twin" is solely determined by "how close" they are in terms of x's.

New Approach

Motivation

- Mass imputation based on CI assumption may not be a good idea.
- The regression of Y_1 on X and Y_2 will provide insignificant regression coefficient on Y_2. That is, the p-value for $\hat{\beta}_2$ will be large in

$$\hat{y}_1 = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 y_2$$

- CI assumption is often unrealistic!
- In many cases, we may have

$$Corr(Y_1, Y_2 \mid X) \neq 0$$

and the true value of β_2 will be different from zero.

Proposal: Kim et al. (2016)

- Data Structure

Table: Data structure

	X	Y_1	Y_2
Sample A	✓	✓	
Sample B	✓		✓

- Model Specification

$$f(Y_1, Y_2 \mid X) = f(Y_1 \mid X; \theta_1) f(Y_2 \mid X, Y_1; \theta_2)$$

- The specified model should be identified under the above data structure.

Remark: Model identification

- Consider the following joint model of (Y_1, Y_2) given X,

$$\begin{aligned} Y_1 &= \alpha_0 + \alpha_1 X + e_1, & (4) \\ Y_2 &= \beta_0 + \beta_1 X + \beta_2 Y_1 + e_2, & (5) \end{aligned}$$

  where e_1 and e_2 are mean zero and $\text{Cov}(e_1, e_2) = 0$. Because (X, Y_1) is observed in sample A, (α_0, α_1) is identifiable. Because (X, Y_2) is observed in sample B, $f(Y_2 \mid X)$ is identifiable.

- Coupling (4) and (5) leads to

$$Y_2 = (\beta_0 + \alpha_0\beta_2) + (\beta_1 + \alpha_1\beta_2)X + \beta_2 e_1 + e_2.$$

  Thus, only $\beta_0 + \alpha_0\beta_2$ and $\beta_1 + \alpha_1\beta_2$ are identifiable and $(\beta_0, \beta_1, \beta_2)$ is not.

- In general, non-linear relationships can help achieve identification. For example, suppose that the linear relationship of X-Y_1 in (4) is changed to

$$Y_1 = \alpha_0 + \alpha_1 X + \alpha_2 X^2 + e_1. \tag{6}$$

- Again, $(\alpha_0, \alpha_1, \alpha_2)$ is identifiable from sample A. Coupling (5) and (6) leads to

$$Y_2 = (\beta_0 + \alpha_0\beta_2) + (\beta_1 + \alpha_1\beta_2)X + (\alpha_2\beta_2)X^2 + \beta_2 e_1 + e_2.$$

Thus, $\beta_0 + \alpha_0\beta_2$, $\beta_1 + \alpha_1\beta_2$ and $\alpha_2\beta_2$ are identifiable from the sample B. As long as $\alpha_2 \neq 0$, $(\beta_0, \beta_1, \beta_2)$ is identifiable.

Suppose that the model is identified. We can apply the EM algorithm for the combined sample.

- E-step: Compute the conditional expectation of the pseudo log-likelihood:

$$
\begin{aligned}
Q_1(\theta_1 \mid \theta^{(t)}) &= \sum_{i \in A \cup B} E\left\{ \log f_1(y_{1i} \mid x_i; \theta_1) \mid x_i, y_{i,obs}; \theta^{(t)} \right\} \\
Q_2(\theta_2 \mid \theta^{(t)}) &= \sum_{i \in A \cup B} E\left\{ \log f_2(y_{2i} \mid x_i, y_{1i}; \theta_2) \mid x_i, y_{i,obs}; \theta^{(t)} \right\}
\end{aligned}
$$

  where $y_{i,obs}$ is the observed part of (y_{1i}, y_{2i}).

- M-step: Update the parameters by finding the maximizer of $Q_1(\theta_1 \mid \theta^{(t)})$ and $Q_2(\theta_2 \mid \theta^{(t)})$ with respect to θ_1 and θ_2.

- Note that

$$
\begin{aligned}
Q_1(\theta_1 \mid \theta^{(t)}) &= \sum_{i \in A} \log f_1(y_{1i} \mid x_i; \theta_1) \\
&+ \sum_{i \in B} E\left\{\log f_1(Y_{1i} \mid x_i; \theta_1) \mid x_i, y_{i2}; \theta^{(t)}\right\} \\
Q_2(\theta_2 \mid \theta^{(t)}) &= \sum_{i \in A} E\left\{\log f_2(Y_{2i} \mid x_i, y_{1i}; \theta_2) \mid x_i, y_{i1}; \theta^{(t)}\right\} \\
&+ \sum_{i \in B} E\left\{\log f_2(y_{2i} \mid x_i, Y_{1i}; \theta_2) \mid x_i, y_{i2}; \theta^{(t)}\right\}
\end{aligned}
$$

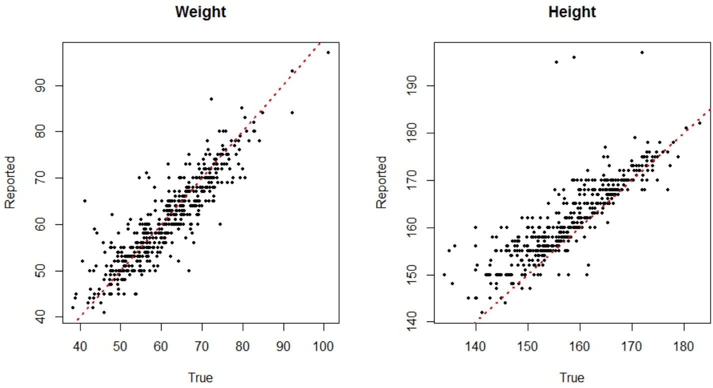
where the conditional expectation in sample A is with respect to $f_2(Y_2 \mid X, Y_1; \theta_2^{(t)})$ and the conditional expectation in sample B is with respect to

$$
f(Y_1 \mid X, Y_2; \theta^{(t)}) = \frac{f_1(Y_1 \mid X; \theta_1^{(t)}) f_2(Y_2 \mid X, Y_1; \theta_2^{(t)})}{\int f_1(Y_1 \mid X; \theta_1^{(t)}) f_2(Y_2 \mid X, Y_1; \theta_2^{(t)}) dY_1}.
$$

4. Measurement error model approach 1: Correcting measurement bias with a validation subsample

Motivating Example: BMI data example

- Korean Longitudinal Study of Aging (KLoSA) data (http://www.kli.re.kr/klosa/en/about/introduce.jsp)
- Original sample measures height and weight from survey questions (N=9,842)
- A validation sample (n=505) is randomly selected from the original sample to obtain physical measurement for the height and weight.


Weight
Height
Reported
True
40
50
60
70
80
90
100
140
150
160
170
180
190

Bayes theorem

- Three random variables
  - X: covariate
  - Y: study variable of interest
  - $\tilde{Y}$: proxy measure of Y with measurement error
- Bayes formula

$$f(y \mid \tilde{y}, x) = \frac{f(\tilde{y} \mid y, x) f(y \mid x)}{\int f(\tilde{y} \mid y) f(y \mid x) d\mu(y)}, \tag{7}$$

where μ is the dominating measure.

- Under the non-differentiability assumption in (1), the above formula can be written as

$$f(y \mid \tilde{y}, x) = \frac{f(\tilde{y} \mid y) f(y \mid x)}{\int f(\tilde{y} \mid y) f(y \mid x) d\mu(y)}. \tag{8}$$

4.1 Model specification

- Two models in (8):
  1. $f(y \mid x) = f_1(y \mid x; \theta)$: process model
  2. $f_2(\tilde{y} \mid y)$: data model

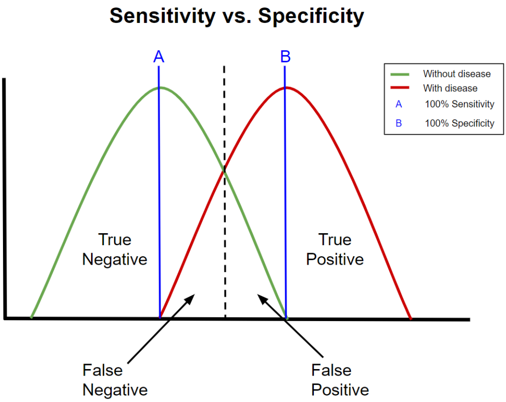
Table: Data structure

	X	Y	$\tilde{Y}$
Sample A		✓	✓
Sample B	✓		✓

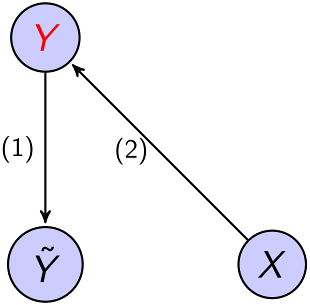
- Sample A can be called calibration sample (or validation sample) as the true measurement y is observed.
- Sample B is the main survey with inaccuarate measurement $\tilde{y}$.
- In some cases, sample A is not available to us. Only the observations in sample B are available. In this case, the data model is treated as known.

Data model for binary Y

- $Y = 1$ means disease status
  - sensitivity (true positive rate): $P(\tilde{Y} = 1 \mid Y = 1) = 1 - \alpha$
  - specificity (true negative rate): $P(\tilde{Y} = 0 \mid Y = 0) = 1 - \beta$



Measurement error model framework



(1): Data model (known),
(2): Process model (known up to θ).

4.2 Parameter estimation: 1. Direct approach

- Idea:
  1. Combine the two models to get the marginal distribution of the observations in sample B

$$
\begin{aligned}
f(\tilde{y} \mid x; \theta) &= \int f_1(y \mid x; \theta) f_2(\tilde{y} \mid y) d\mu(y) \\
&:= \tilde{f}(\tilde{y} \mid x; \theta)
\end{aligned}
$$

  2. Construct the observed log-likelihood function of θ

$$
\ell_{obs}(\theta) = \sum_{i=1}^{n} \log \tilde{f}(\tilde{y}_i \mid x_i; \theta). \tag{9}
$$

  3. Compute the maximizer of $\ell_{obs}(\theta)$:

$$
\hat{\theta} = \arg\max_{\theta} \ell_{obs}(\theta).
$$

4.2 Parameter estimation: 2. EM algorithm

- First define the log-likelihood function using true measurement Y:

$$\ell_{com}(\theta) = \sum_{i=1}^{n} \log f_1(y_i \mid x_i; \theta)$$

- Iterative computation:
  - **E-step**: Given the current parameter $\theta^{(t)}$, compute

$$\begin{aligned} Q(\theta \mid \theta^{(t)}) &= E\{\ell_{com}(\theta) \mid X, \tilde{Y}; \theta^{(t)}\} \\ &= \sum_{i=1}^{n} \left[E\{\log f_1(Y \mid x_i; \theta) \mid x_i, \tilde{y}_i; \theta^{(t)} \right] \end{aligned}$$

    where the expectation is with respect to

$$f(y \mid x, \tilde{y}; \theta^{(t)}) = \frac{f_2(\tilde{y} \mid y) f_1(y \mid x; \theta^{(t)})}{\int f_2(\tilde{y} \mid y) f_1(y \mid x; \theta^{(t)}) d\mu(y)}.$$

  - **M-step**: Update θ by

$$\theta^{(t+1)} = \arg\max Q(\theta \mid \theta^{(t)}). \tag{10}$$

4.3 Prediction (or mass imputation)

- Best prediction: Expectation from the prediction model at $\theta = \hat{\theta}$

$$\hat{Y}_i^* = E\left(Y_i \mid X_i, \tilde{Y}_i; \hat{\theta}\right) \tag{11}$$

  This is a denoised version of $\tilde{Y}_i$ in sample B.

- Prediction model is obtained by combining data model with process model using Bayes theorem:

$$f(y \mid x, \tilde{y}; \theta) = \frac{f_2(\tilde{y} \mid y) f_1(y \mid x; \hat{\theta})}{\int f_2(\tilde{y} \mid y) f_1(y \mid x; \hat{\theta}) d\mu(y)}.$$

4.4 Prediction error

- Let
$$Y_i^* = E\left(Y_i \mid X_i, \tilde{Y}_i; \theta\right) := Y_i^*(\theta).$$
- Prediction error of $\hat{Y}_i^* = Y_i^*(\hat{\theta})$ in (11):
$$\hat{Y}_i^* - Y_i = \{Y_i^*(\theta) - Y_i\} + \left\{Y_i^*(\hat{\theta}) - Y_i^*(\theta)\right\}. \qquad (12)$$
- In (12), the first part is the genuine prediction error and the second part is the error due to the uncertainty in $\hat{\theta}$.
- Mean Squared Prediction Error:
$$\begin{aligned} MSPE(\hat{Y}_i^*) &\doteq E\{(Y_i^* - Y_i)^2\} + B_i V(\hat{\theta}) B_i' \\ &= E\{V(Y_i \mid X_i, \tilde{Y}_i)\} + B_i V(\hat{\theta}) B_i', \end{aligned}$$

where $B_i = \partial Y_i^*(\theta)/\partial\theta$.

Statistical Methods (Summary)

Basic Steps

1. Model Specification
   - Data model
   - Process model
2. Parameter estimation
   - Direct maximization of marginal likelihood
   - EM alghorithm
3. Best prediction
   - Derive the predictive model using Bayes formula
   - Best prediction is obtained by computing the expectation of the prediction model evaluated at MLE.
4. Uncertainty quantification
   - Linearization or Bootstrap
   - Bayesian approach

Discussion

- Parametric fractional imputation of Kim (2011) can be a useful computational tool for the EM algorithm.
- The whole theory and methods for imputation and missing data can be found in Kim and Shao (2021).
- Xu et al. (2017) develop a semiparametric ML estimator using nonparametric estimation of $f_2(\tilde{y} \mid y)$ from the calibration sample.
- Park and Kim (2018) considered a mixture model for $f_2(\tilde{y} \mid y, x)$ and used the model to obtain the prediction for correct BMI in the KLoSA data.

5. Measurement error model approach 2: Correcting measurement bias using external calibration sample

5.1 Introduction

- Data A: Survey sample data
- Data B: Non-survey data
  - Administrative data (Income tax data)
  - Credit card purchase information, voluntary membership database
- Traditionally, only Data A is used for official statistics.
- However, using Data B is becoming more important because
  1. decreasing participation rate in survey samples.
  2. increasing availability of non-survey data

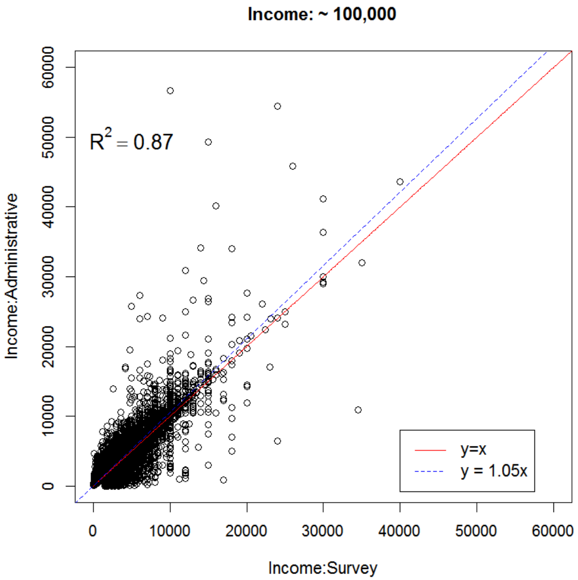
Survey participation rates over time

Economist.com

Motivating Example: Income survey data with Tax information

- Statistics Korea performs a yearly household survey of income and expenditure (with sample size = 20,000).
- From 2015, income tax information is available for the matched sample (about 85%) of the sample.
- Income from survey data is systematically different from Tax income data.
- The non-matched part of the sample can be treated as missing data.

Plot of two income data (wage income)



5.2 Mass imputation under imperfect matching

Table: Data structure for sample A

	X	$\tilde{Y}$	Y
Matched	✓	✓	✓
Unmatched	✓	✓	

- Thus, it is a missing data problem.
- Imputed values are generated from

$$y^* \sim f(y \mid x, \tilde{y}) = \frac{f(y \mid x) f(\tilde{y} \mid x, y)}{\int f(y \mid x) f(\tilde{y} \mid x, y) dy}$$

Example

1. Model Specification:

$$
\begin{aligned}
y_i \mid x_i &\sim f_1(y_i \mid x_i; \theta_1) \\
\tilde{y}_i \mid (x_i, y_i) &\sim f_2(\tilde{y}_i \mid x_i, y_i; \theta_2)
\end{aligned}
$$

2. Parameter estimation: Find the maximizer of

$$
\begin{aligned}
l_{obs}(\theta_1, \theta_2) &= \sum_{i \in A} w_i \delta_i \left\{ \log f_1(y_i \mid x_i; \theta_1) + \log f_2(\tilde{y}_i \mid x_i, y_i; \theta_2) \right\} \\
&+ \sum_{i \in A} w_i (1 - \delta_i) \log \int f_1(y \mid x_i; \theta_1) f_2(\tilde{y}_i \mid x_i, y; \theta_2) dy
\end{aligned}
$$

3. Prediction: Use

$$
y^* \sim \frac{f_1(y \mid x_i; \hat{\theta}_1) f_2(\tilde{y} \mid x_i, y_i; \hat{\theta}_2)}{\int f_1(y \mid x_i; \hat{\theta}_1) f_2(\tilde{y}_i \mid x_i, y; \hat{\theta}_2) dy}
$$

Remark

- Instead of direct maximization of $l_{obs}(\theta)$ for $\theta = (\theta_1, \theta_2)$, one can consider EM algorithm:

  [E-step] Given $\theta^{(t)}$, compute

$$
\begin{aligned}
Q_1(\theta_1 \mid \theta^{(t)}) &= \sum_{i \in A} w_i \delta_i \log f_1(y_i \mid x_i; \theta_1) \\
&\quad + \sum_{i \in A} w_i (1 - \delta_i) E\left\{ \log f_1(Y \mid x_i; \theta_1) \mid x_i, \tilde{y}_i; \theta^{(t)} \right\}
\end{aligned}
$$

$$
\begin{aligned}
Q_2(\theta_2 \mid \theta^{(t)}) &= \sum_{i \in A} w_i \delta_i \log f_2(\tilde{y}_i \mid x_i, y_i; \theta_2) \\
&\quad + \sum_{i \in A} w_i (1 - \delta_i) E\left\{ \log f_2(\tilde{y}_i \mid x_i, Y; \theta_2) \mid x_i, \tilde{y}_i; \theta^{(t)} \right\}
\end{aligned}
$$

  [M-step] Update the parameters by finding the maximizer of Q_1 and Q_2.

2017 KHIES Data

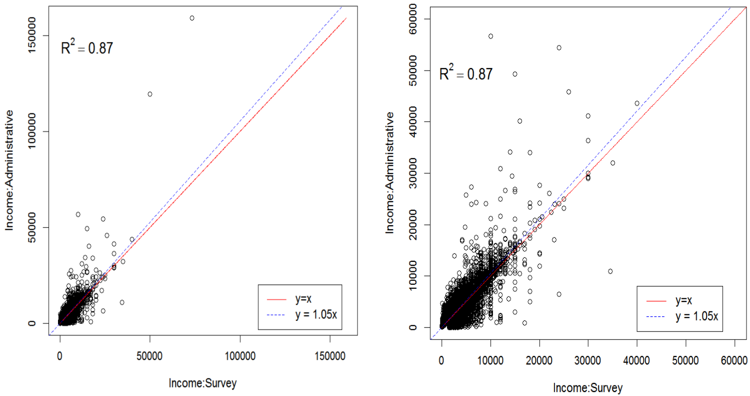
- We apply the proposed method to the 2017 Korean Household Income and Expenditure Survey (KHIES) conducted by Statistics Korea.
- One purpose of the KHIES is to provide up-to-date information about Korean household welfare-related status.
- It measures several different types of income items for each person in a household as well as expenditure-related items and basic demographic information.
  - earned income, business income, financial income, property income, and other types of incomes
- Earned income is the primary variable considered in this study.

5.3 Application to 2017 KHIES Data

- Since 2014, income tax administrative data has been accessible to Statistics Korea.
- The accurate information about earned income is available for each person in the sample using personal identification number (PIN).
- However, some participants in the sample do not reveal PIN. In this case, their tax information about earned income is not available.
- The overall matching rate of the KHIES sample is about 85%.

2017 KHIES Data

Figure: Scatterplots of the survey and administrative earned incomes for the matched respondents in the KHIES (Unit: KRW 10,000)



2017 KHIES Data

Table: Summary statistics of survey and administrative annual earned incomes for the matched and unmatched groups (Unit: KRW 1,000)

		1st Qu.	Median	Mean	3rd Qu.
Matched	Survey	14,400	24,000	31,450	40,000
	Admin.	12,000	22,280	31,990	42,200
Unmatched	Survey	15,000	24,000	29,290	37,100
	Admin.	NA			

- The earned incomes from the two data sources are highly correlated, however, there are still differences, which suggests measurement errors in the reported income in KHIES.

Application to 2017 KHIES

- In Figure 1, we observe that $\tilde{y}$ and y are highly correlated with increasing variation for large $\tilde{y}$
  $\Rightarrow$ Ratio imputation of y using $\tilde{y}$ only is appealing!
- To improve the prediction accuracy, we can divide data into several cells so that observations are homogeneous within each cell and then perform ratio imputations within each cell.
- Such cell-formation can be determined by $\tilde{y}$ and other covariates x.
- However, we do not have clear evidence of a relationship between $\tilde{y}$ and x, and $\tilde{y}$ is very skew-distributed itself.

Gaussian Mixture model

- Decompose Y into (x, y), where y is subject to missingness and x is always observed.
- The GMM is based on the joint model of (x, y)

$$f(x,y) = \sum_{g=1}^{G} P(z=g) f(x,y \mid z=g) = \sum_{g=1}^{G} \pi_g \phi(x,y;\psi_g)$$

  where z is the latent variable taking values on $\{1,2,\ldots,G\}$.
- From the joint distribution, we can derive the conditional distribution of y given x:

$$f(y \mid x) = \sum_{g=1}^{G} P(z=g \mid x)\phi(y \mid x, z=g), \tag{13}$$

  where

$$P(z=g \mid x) = \frac{\pi_g \phi(x \mid z=g)}{\sum_{g=1}^{G} \pi_g \phi(x \mid z=g)}.$$

Conditional Gaussian mixture model (CGMM)

- Under complete responses, instead of deriving (13) from GMM, we directly assume that

$$
\begin{aligned}
f(y \mid x) &= \sum_{g=1}^{G} P(z=g \mid x) f(y \mid x, z=g) \\
&= \sum_{g=1}^{G} \pi_g(x)\phi(y \mid x; \psi_g),
\end{aligned} \tag{14}
$$

  - $\pi_g(x) = P(z=g \mid x)$
  - $\phi(\cdot; \psi_g)$: density of multivariate Gaussian distribution with ψ_g

- Assume that $\pi_g(x) = \pi_g(x; \alpha_g)$ follows a multinomial logit model.

$$
\pi_g(x; \alpha_g) = \frac{exp(\alpha_{g0} + x'\alpha_{g1})}{\sum_{h=1}^{G} exp(\alpha_{h0} + x'\alpha_{h1})} \tag{15}
$$

for $g = 1, \cdots, G$, where $(\alpha_{10}, \alpha_{11}')' = 0_{q+1}$.

Application to 2017 KHIES

- This motivates the CGMM: for $g = 1, \ldots, G$,

$$\begin{aligned} \pi_g(\tilde{y}, x) &= \frac{exp\{(1, \tilde{x}')\alpha_g\}}{1 + \sum_{k=2}^{G} exp\{(1, \tilde{x}')\alpha_k\}}, \\ y_i \mid \tilde{y}_i, x_i, z_i = g &\sim N(\tilde{y}_i \beta_g, \sigma_g^2), \end{aligned}$$

  - $\tilde{x}_i = (\tilde{y}_i, x_i')'$ and $\alpha_1 = 0$

- Imputation model:

$$f(y \mid \tilde{y}, x) = \sum_{g=1}^{G} \pi_g(\tilde{y}, x) f(y \mid \tilde{y}; \beta_g, \sigma_g^2),$$

where

$$\pi_g(\tilde{y}, x) = \frac{exp\{(1, \tilde{x}')\alpha_g\}}{1 + \sum_{k=2}^{G} exp\{(1, \tilde{x}')\alpha_k\}}, \tag{16}$$

Application to 2017 KHIES

- Let $\hat{\theta}$ denote the maximum likelihood estimates and we compute imputed values of y for the unmatched respondents in the survey as

$$\hat{y}_i^* = \sum_{g=1}^{G} \hat{\pi}_g(\tilde{y}_i, x_i)\tilde{y}_i\hat{\beta}_g,$$

  - $\hat{\pi}_g(\tilde{y}, x)$ is $\pi_g(\tilde{y}, x)$ in (16) evaluated at $\alpha = \hat{\alpha}$.
  - a weighted sum of cell ratio estimation.
- We consider $G = \{1, \ldots, 10\}$ and then select G minimizing $BIC(G)$.
  - In this data, $G = 4$ was selected.

Application to 2017 KHIES

Table: Estimated parameters of CGMM with $G = 4$

g	β_g	σ_g^2	$\alpha_{g,0}$	$\alpha_{g,Age}$	$\alpha_{g,Edu}$	$\alpha_{g,Survey}$
1	1.00	0.00	0.00	0.00	0.00	0.00
2	1.03	37.06	0.88	-0.11	-0.08	2.51
3	1.44	5912.03	-1.28	0.38	-0.10	2.63
4	0.96	605.17	1.49	-0.23	-0.05	2.25

- It successfully distinguishes a cell in which the survey and administrative earned incomes are exactly the same, from other cells.
- The survey earned income contributed more to form such cells than age and education.

Application to 2017 KHIES

Table: Summary statistics of survey and administrative/imputed earned incomes for the matched/unmatched groups (Unit: KRW 1,000)

		1st Qu.	Median	Mean	3rd Qu.
Matched	Survey	14,400	24,000	31,450	40,000
	Admin.	12,000	22,280	31,990	42,200
Unmatched	Survey	15,000	24,000	29,290	37,100
	Imputed	15,130	24,310	29,720	37,610

- The average imputed earned income is higher than the mean of the survey earned income:
  - consistent with the difference between the survey and administrative incomes for the matched respondents

Application to 2017 KHIES

Table: Imputation results with 95% confidence interval and estimates of survey earned incomes (Unit: KRW 1,000)

	Survey estimate	Imputed estimate	95% C.I.
1st Qu.	14,450	12,104	(11,904, 12,303)
Median	24,000	22,778	(22,164, 23,391)
Mean	31,204	31,675	(31,213, 32,137)
3rd Qu.	40,000	41,396	(40,592, 42,199)

- The jackknife method is used to estimate the variance of the imputed estimates.
- The proposed imputed results show non-negligible differences from the estimates only based on the survey earned income.
- For more details, see Lee and Kim (2022).

6. Conclusion

- Mass imputation uses missing data imputation to construct synthetic data for unmeasured survey items.
- Fully observed data can be used as training data for model selection and parameter estimation.
- For non-monotone missing data, the EM algorithm can be used with parametric model assumptions.
- Extensions to non-parametric and semi-parametric models need to be developed further.
- Promising area of research and application.

References I

Kim, J. K. (2011). Parametric fractional imputation for missing data analysis. *Biometrika 98*, 119–132.

Kim, J. K., E. Berg, and T. S. Park (2016). Statistical matching using fractional imputation. *Survey Methodology 42*, 19–40.

Kim, J. K. and J. Shao (2021). *Statistical Methods for Handling Incomplete Data* (2nd ed.). Chapman & Hall / CRC.

Lee, D. and J. K. Kim (2022). Semiparametric imputation using conditional Gaussian mixture models under item nonresponse. *Biometrics 78*, 227–237.

Park, S. and J. K. Kim (2018). Analysis of inaccurate data using mixture measurement error models. *Journal of the Korean Statistical Society 47*, 1–12.

Xu, Y., J. K. Kim, and Y. Li (2017). Semiparametric estimation for measurement error models with validation data. *Canadian Journal of Statistics 45*, 185–201.