Data Integration for Split Questionnaire

Jongho Im, Sanghyo Kim and Inho Park

2021 춘계통계학회

May 28, 2021

Outline

1. Introduction
2. Fractional Imputation
3. Simulation Study
4. Real data analysis
5. Concluding Remarks

Introduction: motivation

Korea Rural Economic Institute (KREI) has been conducted two annual surveys from 2018:

1. Consumer Behavior Survey for Foods (CBSF).
2. Consumer Attitude Survey for Processed Foods (CASPF).

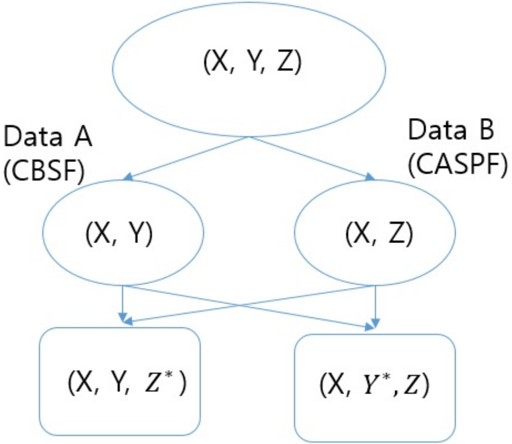
Long-term goal: Split questionnaire survey

Short-term goal: Analysis combining two surveys in micro-level.

For example, interested in $Y \mid Z$:

- (CBSF) Y: Changes in HMR consumption relative to last year.
- (CASPF) Z
  - frequency of dining-out.
  - evaluation on food safety of HMRs (Home Meal Replacements).

Introduction: data structure



Data structure

- X are commonly observed on both A and B.

Introduction: goals

Under our basic setup, interested in

$$(Y_i, X_i) \longrightarrow (Y_i, X_i, Z_i^*) \ \ i \in A,$$

Suffices to know the distribution of Z conditional on Y and X,

$$f(Z \mid Y, X) \propto f_1(Y \mid X, Z) f_2(Z \mid X)$$

Or, to know the conditional distribution of f_1 when f_2 can be correctly estimated from Data B.

Introduction: some poplar methods

1. Conditional independent assumption

$$f(Z \mid Y, X) = f_2^*(Z \mid X)$$

   - f_2^* can be estimated from Data B.
   - Weired if we are interested in a regression of $Y \mid X, Z$.

2. Record Linkage

   - Often there exists common units between two data sets.
   - Require key variables in X such as identification variables or demographic variables.

3. Nearest Neighbor Imputation

   - Find donors using commonly observed X values.
   - Hard to represent the relationship between Y and Z given X.

Fractional Imputation: short review

- Initially proposed by Kalton and Kish (1984) and extensively discussed in Fay (1996).
- Kim and Fuller (2004) and Fuller and Kim (2005) proposed fractional hot deck imputation (FHDI) as a repeated imputation.
- Kim (2011) proposed a parametric fractional as a general tool for missing data analysis. (MCEM)
- Im, Kim, Fuller (2015) investigated a multivariate version of FHDI and the R package 'FHDI' was developed by Im, Cho and Kim (2018).

Fractional Imputation: basic idea

- $E(y_{i,mis} \mid y_{i,obs})$ is approximated by

$$E(y_{i,mis} \mid y_{i,obs}) \cong \sum_{j=1}^{M} w_{ij}^{*} y_i^{*(j)},$$

  - $(y_{i,mis}, y_{i,obs})$ is the (observed, missing) part of y_i.
  - M: a size of imputed values on the unit i
  - $y_i^{*(j)}$: j-th imputed value for $y_{i,mis}$, $j = 1, \ldots, M$.
  - w_{ij}^{*}: fractional weight assigned to the j-th imputed value (vector) of unit i.

- Split the record with missing item into $M(> 1)$ imputed values.
- Assign fractional weights on imputed values.
- The final product is a single data file with size $\leq nM$.
- For variance estimation, the fractional weights are replicated.

Fully Nonparametric FHDI 1

Fully Nonparametric FHDI

1. Both f_1 and f_2 are estimated non-parametrically.
2. Relatively easy to extend for multivariate variables.

Assumption

- All variables are categorical. (Note: can apply categorization for interval data in practice).
- Some of X are instrumental variables for Z.

Fully Nonparametric FHDI 2

(Step 1) Estimate joint cell probabilities $P(X, Z)$ and $P(X_2, Z)$ from data B, where $X = (X_1, X_2)$.

(Step 2) Impute all possible Z values for each observation in data A.

Table: An illustrative example with binary responses

ID	Y	X_1	X_2	Z^*
1	1	1	1	1^*
1	1	1	1	2^*
2	2	2	1	1^*
2	2	2	2	2^*

Fully Nonparametric FHDI 3

(Step 3) Compute fractional weights w_{ij}^*.

E-step

$$
\begin{aligned}
w_{ij}^* &\propto P(Y \mid X, Z^*)P(Z^* \mid X) \\
&\propto \frac{P(Y, X, Z^*)}{P(Z^*, X)} P(Z^*, X) \\
&= \frac{P(Y, X_2, Z^*)}{P(Z^*, X_2)} P(Z^*, X) \quad \text{(Asumption)}
\end{aligned}
$$

- $P(Z^*, X)$ and $P(Z^*, X_2)$ were estimated on data B (Step 1).
- $P(Y, X_2, Z^*)$ is estimated on imputed data A (M-step).
- X_1 in $X = (X_1, X_2)$ plays a role of instrumental variable for identification.
- w_{ij}^* are normalized so that $\sum_j w_{ij}^* = 1$ for all i's.

Fully Nonparametric FHDI 4

(Step 3) Compute fractional weights w_{ij}^*.

M-step: update joint cell probabilities of $P(Y, X, Z^*)$

$$P(Y = y, X = x, Z^* = z) = \frac{\sum_{i,j} w_i w_{ij}^* I(Y_i = y, X_i = x, Z_i^{*(j)} = z)}{\sum_{i,j} w_i w_{ij}^*}.$$

(Step 4) Select M imputed values for each recipient $i \in A$ with the probability proportional to w_{ij}^*.

Note that the current algorithm allows missing values on both data A and data B.

Simulation Study 1

- X_1, X_2 and Z are separately generated from Bernoulli distribution ($p = 0.5$), but linked through a Gaussian copula with the correlation matrix.

$$\begin{pmatrix} 1 & 0.3 & 0.5 \\ 0.3 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{pmatrix}$$

- Y is generated from a logit model

$$P(Y = 1 \mid X_1, X_2, Z) = \{1 + \exp(-\beta_0 - \beta_1 X_2 - \beta_2 Z)\}^{-1},$$

  where $\beta_0 = 0$, $\beta_1 = 0.5$ and $\beta_2 = 1$.
- $n = 2,000$ samples with $n_A = 1,000$ and $n_B = 1,000$: (Y, X_1, X_2) in data A and (X_1, X_2, Z) for data B.
- Working model: $Y \mid (X_2, Z)$.

Simulation Study 2

Table: MC results for simulation study 1

	Full		NNI		FHDI		
	EST	SE	EST	SE	EST	SE	SE2
β_1	0.50	0.16	0.87	0.16	0.50	0.28	0.16
β_2	1.01	0.18	-0.01	0.17	1.00	0.45	0.18

- Nearest neighbor was found in comparison X_1 and X_2.
- Single value $M = 1$ is generated for each Z_i.
- EST denotes point estimates.
- SE denotes MC variance and SE2 denotes the estimated SE as if the imputed values were originally observed.

Real Data analysis 1

1. 2019년 '식품소비 행태조사'와 '가공식품 소비자 태도조사'의 연계분석 가능성을 검정하기 위하여 총 1,519 가구에 대하여 추가 조사가 이뤄졌음.

2. 성별, 연령대, 식품첨가물 인식, MSG 인식, 간편식 이용 여부가 모두 관측되는 validation sample이 존재.

변수 구분	변수명	내용
공통변수	X_1	성별(남성/여성)
	X_2	연령대(20-30/40-50/60+)
가공식품	Y_1	식품첨가물인식 (상관없음/먹지않음)
소비자태도조사	Y_2	MSG인식(상관없음/먹지않음)
식품소비행태조사	Z	간편식이용 여부 (예/아니오)

Real Data analysis 2

(X_2, Y_1, Y_2, Z)	Full	CI	NNI	FHDI
(1,1,1,1)	0.053	0.048	0.049	0.049
(1,1,1,2)	0.007	0.048	0.011	0.011
(1,1,2,1)	0.026	0.026	0.025	0.026
⋮	⋮	⋮	⋮	⋮
(3,2,1,2)	0.030	0.027	0.27	0.023
(3,2,2,1)	0.027	0.040	0.38	0.035
(3,2,2,2)	0.049	0.040	0.38	0.041

- Full-validation sample; CI-Conditional Independence; NNI-Nearest Neighbor imputation; FI-fractional hot deck imputation
- RMSE: CI-0.0233, NNI:0.0031, FHDI:0.0021.
- FHDI additionally represents the relationship between Y and Z.

Concluding Remarks

- FHDI can be used to handle split questionnaires (intended non-response).

- When the items are not so large, the proposed method works well.

- For the large items, we need more entity observations. It would be appropriate for big data analysis.

- For non-survey data, we first need to adjust selection bias if we have benchmark information.

$$\begin{aligned} \sum_{i \in A} w_i x_i &= X \\ \sum_{j \in B} w_j x_j &= X \end{aligned}$$

- Identification issue is under development as the future study.