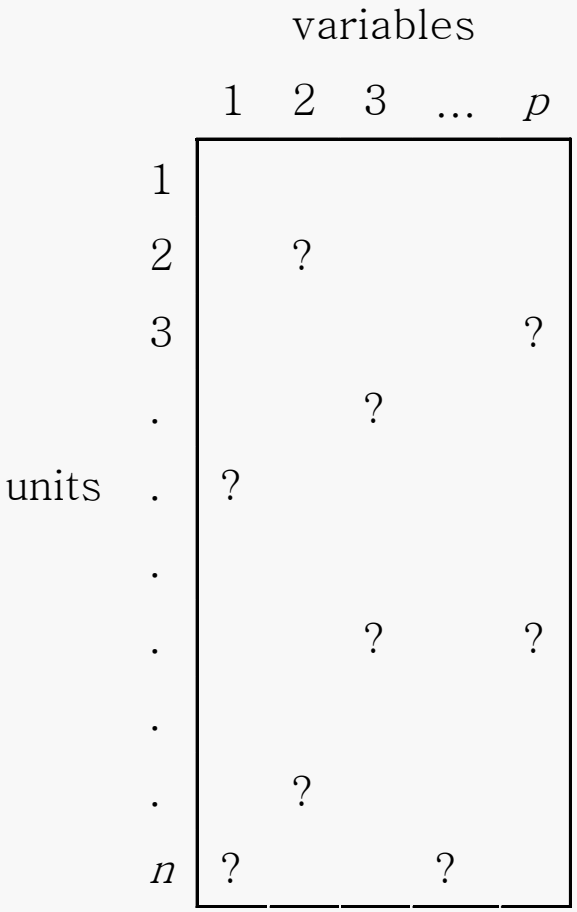
Multiple Imputation for Missing Data in KLoSA

Juwon Song

Korea University and UCLA

Contents

Typical Dataset with Missing Values



Missing Data Mechanisms

- Notation

- $Y = (y_{ij})$: $(n \times p)$ data set

  Y_{obs}: the observed components of Y

  Y_{mis}: the unobserved (missing) components of Y

- Missing-data indicator matrix $M = (m_{ij})$ such that

  $m_{ij} = 1$ if y_{ij} is missing

  $m_{ij} = 0$ if y_{ij} is observed

- $f(Y|\theta) = f(Y_{obs}, Y_{mis}|\theta)$: joint distribution of Y_{obs} and Y_{mis},
  where θ indicates unknown parameters.

- $f(M|Y, \phi)$: conditional distribution of M given Y ,
  where ϕ indicates unknown parameters.

Missing Data Mechanisms

- Full model treats M as a random variable and specifies the joint distribution of M and Y :

$$f(Y, M \mid \theta, \phi) = f(Y \mid \theta)\, f(M \mid Y, \phi), \quad \text{for } (\theta, \phi) \in \Omega_{\theta,\phi},$$

where $\Omega_{\theta,\phi}$ is the parameter space of (θ, ϕ).

- Observed data model

$$f(Y_{obs}, M \mid \theta, \phi) = \int f(Y, M \mid \theta, \phi)\, dY_{mis}$$
$$= \int f(Y_{obs}, Y_{mis} \mid \theta)\, f(M \mid Y_{obs}, Y_{mis}, \phi)\, dY_{mis}.$$

- The likelihood of θ and ψ

$$L(\theta, \phi \mid Y_{obs}, M) \propto f(Y_{obs}, M \mid \theta, \phi)$$
$$= \int f(Y_{obs}, Y_{mis} \mid \theta)\, f(M \mid Y_{obs}, Y_{mis}, \phi)\, dY_{mis}.$$

Missing Data Mechanisms

- MCAR (Missing Completely At Random)
  - $f(M|Y, \phi) = f(M|\phi)$ for all Y, ϕ
  - Missing items are a random subsample of all data values.

- MAR (Missing At Random)
  - $f(M|Y, \phi) = f(M|Y_{obs}, \phi)$ for all Y_{mis}, ϕ
  - The probability that an observation is missing may depend on observed quantities but not on unobserved quantities.

- NMAR (Not Missing At Random)
  - The mechanism is called NMAR if the distribution of M depends on the missing values in the data matrix Y.

- Ignorable
  - When the missing data mechanism is either MCAR or MAR, and the parameters of data and the parameters of the missing data mechanism are distinct.

Imputation

- Imputation: methods to impute the values of items that are missing.

- Imputation based on explicit modeling

  - The predictive distribution is based on a formal statistical model.
  - The assumptions are explicit.
  - Ex) Unconditional mean imputation
    Conditional mean imputation
    Probability imputation
    Regression imputation
    Stochastic regression imputation
    Imputation based on multivariate normal distribution
    Imputation based on nonnormal distributions

Imputation

- Imputation based on implicit modeling

  - The focus in on an algorithm, which implies an underlying model.
  - The assumptions are implicit.
  - Ex) Hotdeck imputation
    Colddeck imputation

- Composite methods are also possible.

  - Ex) Hotdeck imputation based on predictive mean matching

Single Imputation

- Single imputation: impute one value for each missing item.

- Problems of single imputation

  - Imputing a single value for a missing value treats the imputed value as known.

  - Without special adjustments, inferences about parameters based on the filled-in data do not account for imputation uncertainty.

  - Standard errors computed from the filled-in data are systematically underestimated.

Variance Estimation Under Single Imputation

- Conduct single imputation and obtain unbiased or nearly unbiased variance estimators:

(1) Derive theoretically an approximate variance formula for the given estimator of interest.

(2) Use the replication methods, which create a number of replicated datasets (called pseudo-replicates) and estimates the variance of a given estimator by the sample variance of replicate estimators.

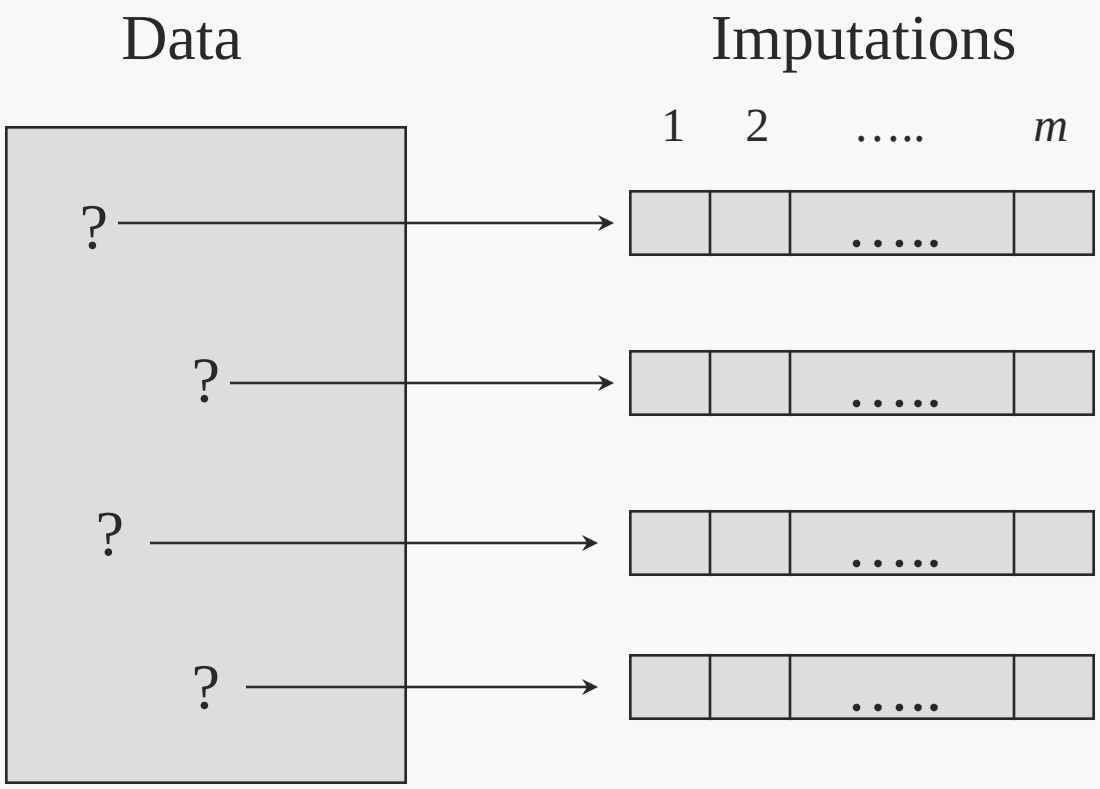
Multiple Imputation

- Multiple Imputation: Impute $m \geq 2$ plausible values for each missing item.

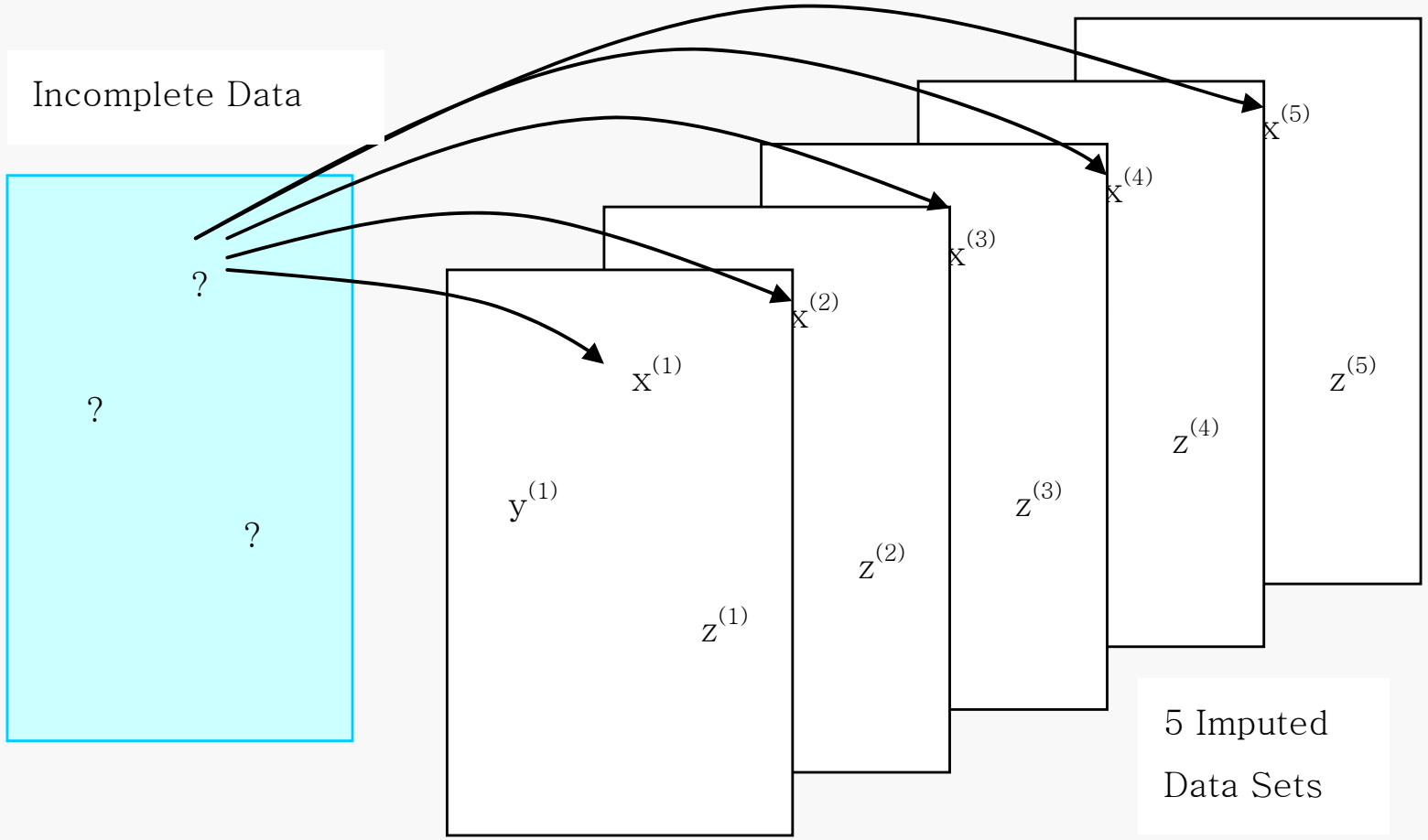
- Generate m complete sets of data.
- Variability among m imputed values provides uncertainty due to missing values.
- Use standard complete-case analysis method for each imputed data and combine the results for the inference.
- Disadvantage over single imputation: more work to create the imputations and analyze the results.

- Many popular multiple imputation models assume that missing data mechanism are MAR.

Multiple Imputation



Example: 5 Multiply Imputed Data Sets



Missing Data in KLoSA

- Korean Longitudinal Study of Aging (KLoSA)

  - Purpose: (1) Evaluate aging trends in the Korean population,
    and (2) apply the findings to the social welfare and labor policy.
  - Sampled 10,254 Koreans aged over 45 from 6,171 families.
  - Longitudinal study: Baseline in 2006
    1st follow-up in 2008
    2nd follow-up in 2010

- As most survey data, KLoSA include missing values.

  - Complete-case analysis may be biased estimates under MAR, and inefficient.
  - Major outcome variables (income and asset related variables) often include missing values.

Missing data in Baseline KLoSA

- Percentage of Missing Values

  - Most variables: < 5%

  - Some Income and asset variables: 10-20%, up to 30%

Session	VARIABLE	N OBS	N MISS	MISSING %
Demographic	Gender	10254	0	0
	Age	10254	0	0
	Educational level	10254	7	0.07
	Marital status	10254	2	0.0002
	Religion	10254	0	0
	Number of family members	10254	0	0
	Number of generations in a family	10254	0	0
Design	Geographic Region	10254	0	0
	Urban/ Rural	10254	0	0
	Housing type	10254	0	0
Income	Wage Income	1986	124	6.24
	Income from own business	1513	97	6.41
	Earning from agricultural/fisheries business	817	24	2.94
	Earning from side job	159	5	3.14
	Total household income	10254	869	8.47
Asset	House market price	7811	1170	14.98
	Total financial asset	4277	682	15.95

Multiple Imputation in Baseline KLoSA

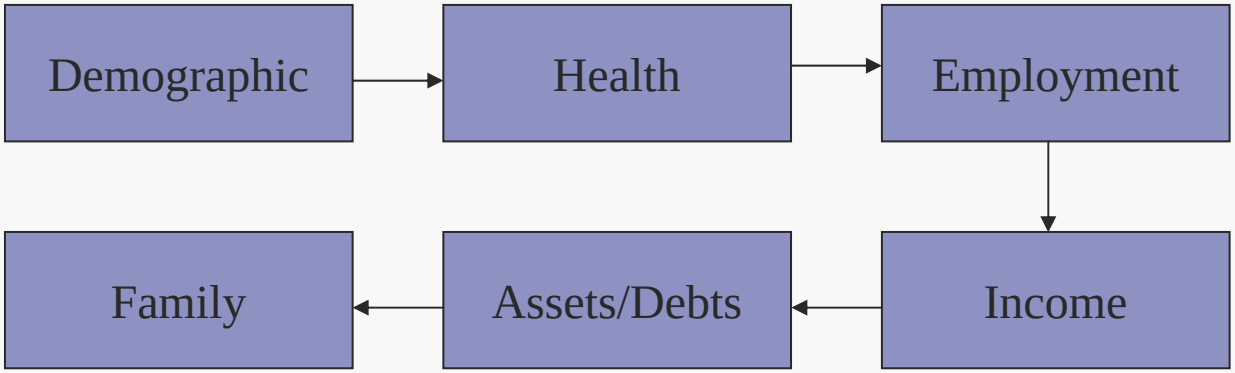
- Questionnaire: consisted of 8 sections

- Cover screen
- Demographic
- Family and family transfer : family representative
- Health
- Employment
- Income
- Assets and debts
- Expectations and life satisfaction session

Multiple Imputation in Baseline KLoSA

- Multiple Imputation

- Focused on income and asset variables.
- Conducted sequentially session by session.



- Five sets of imputed values: Allows variability due to imputation.
- A multiple imputation method was chosen after a simulation of major variables.
- Chosen imputation method: Hotdeck based on a predictive mean matching

Characteristics of Income and Asset Variables

- Use of unfolding brackets

- Include unfolding bracket questions to obtain at least partial information about missing or inconsistent income and asset values.

E005. Did it amount to a total of less than, about equal to or more than 600MW(10,000won)?
[1] Less than 600MW
[3] About 600MW
[5] More than 600MW

E006. Did it amount to a total of less than, about equal to, or more than 1,200MW(10,000won)?
[1] Less than 1,200MW
[3] About 1,200MW
[5] More than 1,200MW

E007. Did it amount to a total of less than, about equal to or more than 2,400MW(10,000won)?
[1] Less than 2,400MW
[3] About 2,400MW
[5] More than 2,400MW

E008. Did it amount to a total of less than, about equal to or more than 6,000MW(10,000won)?
[1] Less than 6,000MW
[3] About 6,000MW
[5] More than 6,000MW

E009. Did it amount to a total of less than, about equal to, or more than 12,000MW(10,000won)?
[1] Less than 12,000MW
[3] About 12,000MW
[5] More than 12,000MW

Table 2. Frequency of the exact answer, unfolding bracket answer and nonresponse for selected questions

SESSION	VARIABLE	Total N	Exact	Range	Missing
Income	Wage Income	1986	1862	82	42
	Income from own business	1513	1416	51	46
	Earning from agricultural/fisheries business	817	793	10	14
	Earning from side job	159	154	5	0
	Total household income	10254	9385	515	354
Asset	House market price	7811	6641	906	264
	Total financial asset	4277	3595	646	36

Characteristics of Income and Asset Variables

- Use of unfolding brackets

  - When additional information were obtained using unfolding brackets, they were measured as ranges.
  - Should incorporate information obtained from unfolding bracket questions to conduct imputation of the exact value.

- Maintaining consistency among variables

  - Some variables in questionnaire are related to each other.
  - Imputation should maintain consistency among variables.

- Several possible imputation methods were considered.

Random Hotdeck Imputation

- Random hotdeck

- In hotdeck imputation, missing values are replaced by recorded values of data.
- Imputed data are in the appropriate range, since they were imputed from other observed values.
- For participants who answered for unfolding bracket questions, missing values are replaced by recorded values from the same unfolding bracket.
- A problem of hotdeck using unfolding brackets is that there may be not many observed participants in some brackets, especially at the top-open bracket.
- Suggested a mixed approach to combine Hotdeck imputation with regression imputation for top-open brackets.
- Adopted for Health and Retirement Study(HRS) in U.S.
- Program: IMPUTE (SAS Macro)

Hotdeck Imputation Based on Predictive Mean Matching

- Hotdeck multiple imputation procedure that used a predicted mean matching method (Little 1998)

- Cycling through each missing-data pattern on each variable with incomplete items, this is consisted of the two-steps:

  (1) forming imputation classes based on the predicted mean of the variable being imputed from a multiple regression model,

  (2) drawing imputations at random from observed data within each class based on an approximate Bayesian bootstrap (ABB).

- For participants who answered for unfolding bracket questions, missing values are replaced by recorded values from the same unfolding bracket.

- Used a mixed approach to combine Hotdeck imputation with regression imputation for top-open brackets.

- Program: SAS MACRO

Sequential Regression Multiple Imputation

- Multiple imputation using a sequence of regression models (Raghunathan et al., 2001)

- Allow imputation using various distributions appropriate to each variable.
- Avoid difficulty of building a full Bayesian models for various types of variables with a sequence of simple multiple regression imputations.
- Model each variable with a conditional density through an appropriate regression model given other variables.

Type of Variables	Model
Continuous	Normal linear regression model
Binary	Logistic regression model
Categorical	Polytomous or generalized logit regression model
Count	Poisson loglinear model
Mixed	Two-stage model

- Conduct multiple imputation using an iterative scheme among conditional distributions.

Sequential Regression Multiple Imputation

- Target joint density to draw

$$f\left(Y_1, Y_2, \ldots, Y_p \middle| X, \theta_1, \theta_2, \ldots, \theta_p\right)$$
$$= f\left(Y_1 \middle| X, \theta_1\right) f\left(Y_2 \middle| X, Y_1, \theta_2\right) \cdots f\left(Y_p \middle| X, Y_1, Y_2, \ldots, Y_{p-1}, \theta_p\right)$$

- Instead, use an approximation by the conditional density:

For the (t +1) iteration, draw

$$f\left(Y_j \middle| X, Y_1^{(t+1)}, Y_2^{(t+1)}, \ldots, Y_{j-1}^{(t+1)}, Y_{j+1}^{(t)}, \ldots, Y_p^{(t)}, \varphi_p\right)$$

- Improve the approximation using the SIR algorithm.

- Multiple imputation using a sequence of regression models
  - Can handle values with limited range.
  - Can handle data collected from sampling strata.
  - Program: IVEWARE (SAS MACRO)

Simulation

- Simulation data

- Considered initial respondents of the KLoSA baseline survey as a population.
- Drew a simple random sample of 250 individuals from male and 250 from female.
- Fitted a logistic model to predict the probability of occurring missing values.
- Individuals were divided as four groups by the predictive probabilities in each gender and 10% of them were considered as missing as follows:

  (1) In the lowest group, 5% of individuals were imposed as missing.

  (2) In the second lowest group, 3% of individuals were imposed as missing.

  (3) In the third lowest group, 2% of individuals were imposed as missing.

  (4) In the highest group, no one was imposed as missing.
- Values corresponding to the missing individuals were changed into unfolding bracket information.

Simulation

- Hotdeck imputation based on a predictive mean matching was compared with other imputation methods using a simulation study.

- Imputation methods

  - Random hotdeck multiple imputation
  - Hotdeck multiple imputation based on a predictive mean matching (chosen)
  - Sequential regression multiple imputation
  - Median imputation
  - Complete-case analysis

- The simulation was conducted for major income/asset variables.

  - Impose missingness using missing percentage of KLoSA baseline data under the MAR mechanism.

Simulation

<Table 1> Wage Income (w01E003)

	Mean	SE	Bias	MSE	Coverage
All data	2184.9	79.36			
Complete-case	2144.3	80.37	-50.93	2432.03	100
Median	2148.47	77.40	-47.19	1702.76	100
IVEWARE	2188.79	83.08	2.94	584.14	100
Hotdeck	2183.65	79.91	-1.74	188.50	100
IMPUTE	2184.53	79.38	-0.55	173.89	100

<Table 2> House Market Price (w01f005)

	Mean	SE	Bias	MSE	Coverage
All data	14422.21	581.48			
Complete-case	14215.71	607.85	-36.09	137907.1	99.1
Median	14643.87	575.60	38.68	66328.33	100
IVEWARE	14637.21	575.6	36.56	73151.79	100
Hotdeck	14381.52	589.22	-7.48	17752.65	100
IMPUTE	13098.02	417.95	-317.89	1919373.92	10.7

Multiple Imputation in Baseline KLoSA Data

- Modified hotdeck imputation using the predictive mean matching to handle various types of variables with missing values.

  - For categorical variables, predictive mean was calculated based on the generalized linear model.

- Extended hotdeck imputation using predictive mean matching.

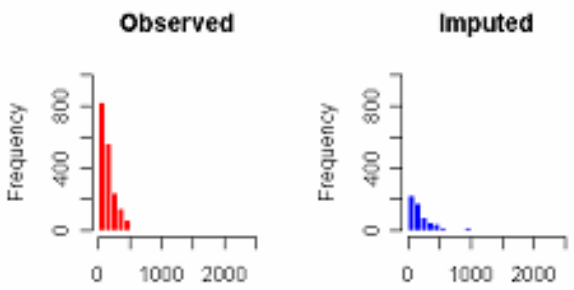
  - Handle unfolding brackets.
  - Work when there are not enough donors within some adjustment cells.
  - Maintain consistency among variables.
  - Incorporate dependency among family members.

- Imputation was conducted separately for male and female.

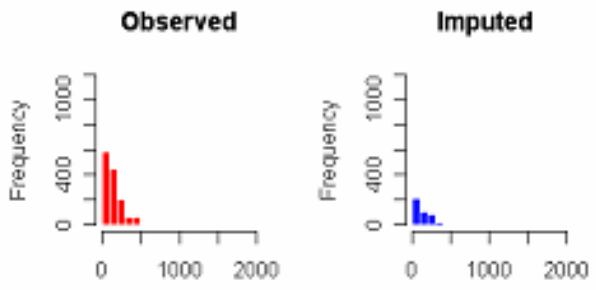
  - Income and asset variables have different distributions between male and female.
  - Covariates in the regression model were chosen among variables that are related to both the response variable and missingness.

Multiple Imputation in Baseline KLoSA Data

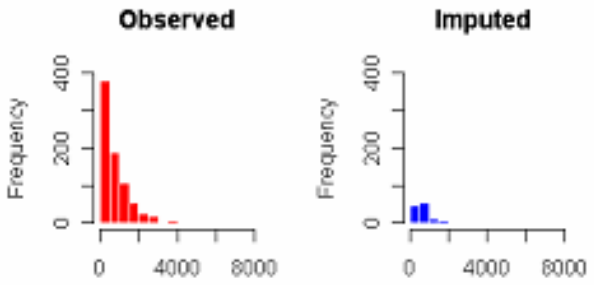
(1) Wage Income



(2) Income from own business



(3) Earning from agricultural/fisheries



(4) Earning from side job

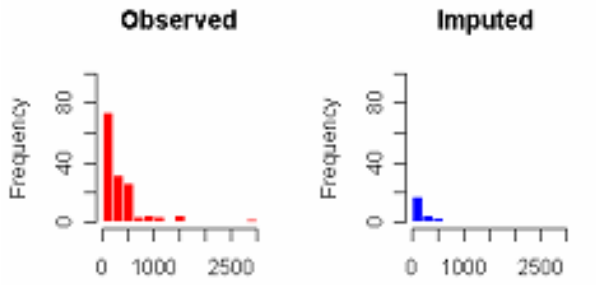


Table 3. Comparison of summary statistics between observed values and imputed values

SESSION	VARIABLE		Min.	1st Q	Med	Mean	3rd Q	Max
Income	Wage Income	Observed	1	80	120	171	220	1000
		Imputed	1	90	150	241	300	2040
	Income from own business	Observed	1	100	150	208	250	10000
		Imputed	50	71	150	244	300	2000
	Earning from agricultural/fisheries business	Observed	1	300	600	942	1200	8000
		Imputed	30	300	600	780	1000	7200
	Earning from side job	Observed	1	100	240	395	500	3000
		Imputed	30	70	120	202	300	600
	Total household income	Observed	0	0.3	1.2	1.9	3	60
		Imputed	0	0.5	1.8	2.3	3	60
Asset	House market price	Observed	0.1	5	10	15.4	16	300
		Imputed	0.1	5	10	17.4	20	300
	Total financial asset	Observed	50	100	300	900	1000	50000
		Imputed	50	200	600	1666	1000	50000

Missing Data in 1st follow-up KLoSA Data

- 1st follow-up KLoSA data

- Include both unit and item missing values.
- Unit nonresponses were handled by weighting methods.
- Item nonresponses were handled by multiple imputation.

- Hotdeck imputation based on the predictive mean matching was chosen to be consistent with imputation of baseline data.

- Since baseline values of a variable are highly correlated with follow-up values of the one, the imputation model included the baseline values as covariates.

Table 1. Percentage of missing values

SESSION	VARIABLE	N OBS	MISSING N	MISSING %
Income	Wage income	1640	31	1.89
	Income from own business	1604	12	0.75
	Earning from agricultural/fisheries business	773	9	1.16
	Earning from side job	45	3	6.67
	Total household income	8688	174	2.00
Asset	House market price	7040	271	3.85
	Total financial asset	538	7	1.30

Multiple Imputation in 1st Follow-up KLoSA Data

Discussion

- Missing data usually occur in survey data.

- Imputation is a popular technique to handle missing data.
  - Both explicit modeling and Implicit one have advantages and disadvantages.
  - Choosing the best imputation model is important.
  - Simulation is useful to choose the imputation model.

- Multiple imputation for the KLoSA study
  - Extended hotdeck imputation to handle unfolding brackets.
  - Modified it to incorporate regression imputation when there were not enough donors in some brackets.
  - Adopted imputation to reserve consistency among variables.
  - Incorporated dependency among family members.

Discussion

- Imputation of Family session

  - Asks financial support from and to each family member, resulting in multiple responses.
  - Incorporate dependency of financial support among family members.
  - The predictive mean in the imputation model was calculated by GEE.
  - Hotdeck imputation based on multilevel modeling (Yoon, 2010)

- Hotdeck Imputation of categorical variables

  - The predictive mean is not easy to define for variables with nominal categories.
  - May be handled similarly to multiple variable cases.

- Imputation of approximate values in unfolding bracket questions

  - How to handle approximate answers is worthy to pursue.