Fractional Imputation Method for Missing Data Analysis in Survey Sampling: A Review

Shu Yang Jae-Kwang Kim

Iowa State University

October 17, 2014

Outline

- Introduction
- Fractional imputation
- Simulation study
- Conclusion

Introduction

Basic Setup

- $U = \{1, 2, \cdots, N\}$: Finite population
- $A \subset U$: sample (selected by a probability sampling design).
- The parameter of interest, $\eta_g = N^{-1} \sum_{i=1}^{N} g(y_i)$. Here, $g(\cdot)$ is a known function.
- For example, $g(y) = I(y < 3)$ leads to $\eta_g = P(Y < 3)$.
- Under complete response, suppose that

$$\hat{\eta}_{n,g} = \sum_{i \in A} w_i g(y_i)$$

is an unbiased estimator of η_g

- What if some of y_i are not observed ?
- Imputation: Fill in missing values by a plausible value (or by a set of plausible values)
- Why imputation ?
  - It provides a complete data file: we can apply the standard complete data methods
  - By filling in missing values, the analyses by different users will be consistent.
  - By a proper choice of imputation model, we may reduce the nonresponse bias.
  - Retain records with partial information: Makes full use of information. (i.e. reduce the variance)

Introduction

Basic Setup (Cont'd)

- $A = A_R \cup A_M$, where y_i are observed in A_R. y_i are missing in A_M
- $\delta_i = 1$ if $i \in A_R$ and $\delta_i = 0$ if $i \in A_M$.
- y_i^*: imputed value for y_i, $i \in A_M$
- Imputed estimator of η_g

$$\hat{\eta}_{I,g} = \sum_{i \in A_R} w_i g(y_i) + \sum_{i \in A_M} w_i g(y_i^*)$$

- Need $E\{g(y_i^*) \mid \delta_i = 0\} = E\{g(y_i) \mid \delta_i = 0\}$.

- Often, find x (always observed) such that
  - Missing at random (MAR) holds: $f(y \mid x, \delta = 0) = f(y \mid x)$
  - Imputed values are created from $f(y \mid x)$.
- An unbiased estimator of η_g under MAR:

$$\hat{\eta}_g = \sum_{i \in A_R} w_i g(y_i) + \sum_{i \in A_M} w_i E\{g(y_i)|x_i\}$$

- Computing the conditional expectation can be a challenging problem.
  1. Do not know the true parameter θ in $f(y \mid x) = f(y \mid x; \theta)$:

$$E\{g(y) \mid x\} = E\{g(y_i) \mid x_i; \theta\}.$$

  2. Even if we know θ, computing the conditional expectation can be numerically difficult.

- Imputation: Monte Carlo approximation of the conditional expectation (given the observed data).

$$E\left\{g\left(y_i\right) \mid x_i\right\} \cong \frac{1}{M}\sum_{j=1}^{M} g\left(y_i^{*(j)}\right)$$

1. Bayesian approach: generate y_i^* from

$$f\left(y_i \mid x_i, y_{obs}\right) = \int f\left(y_i \mid x_i, \theta\right) p(\theta \mid x_i, y_{obs}) d\theta$$

2. Frequentist approach: generate y_i^* from $f\left(y_i \mid x_i; \hat{\theta}\right)$, where $\hat{\theta}$ is a consistent estimator.

- Thus, imputation is a computational tool for computing the conditional expectation $E\{g(y_i) \mid x_i\}$ for missing unit i.
- To compute the conditional expectation, we need to specify a model $f(y \mid x; \theta)$ evaluated at $\theta = \hat{\theta}$.
- Thus, we can write $\hat{\eta}_{I,g} = \hat{\eta}_{I,g}(\hat{\theta})$.
- To estimate the variance of $\hat{\eta}_{I,g}$, we need to take into account of the sampling variability of $\hat{\theta}$ in $\hat{\eta}_{I,g} = \hat{\theta}_{I,g}(\hat{\theta})$.

Three approaches

- Bayesian approach: multiple imputation by Rubin (1978, 1987), Rubin and Schenker (1986), etc.
- Resampling approach: Rao and Shao (1992), Efron (1994), Rao and Sitter (1995), Shao and Sitter (1996), Kim and Fuller (2004), Fuller and Kim (2005).
- Linearization approach: Clayton et al (1998), Shao and Steel (1999), Robins and Wang (2000), Kim and Rao (2009).

Comparison

	Bayesian	Frequentist
Model	Posterior distribution $f(\text{latent}, \theta \mid \text{data})$	Prediction model $f(\text{latent} \mid \text{data}, \theta)$
Computation	Data augmentation	EM algorithm
Prediction	I-step	E-step
Parameter update	P-step	M-step
Parameter est'n	Posterior mode	ML estimation
Imputation	Multiple imputation	Fractional imputation
Variance estimation	Rubin's formula	Linearization or Bootstrap

Fractional Imputation

Idea (parametric model approach)

- Approximate $E\{g(y_i) \mid x_i\}$ by

$$E\{g(y_i) \mid x_i\} \cong \sum_{j=1}^{M_i} w_{ij}^* g(y_i^{*(j)})$$

where w_{ij}^* is the fractional weight assigned to the j-th imputed value of y_i.

- If y_i is a categorical variable, we can use

$$\begin{aligned} y_i^{*(j)} &= \text{the } j\text{-th possible value of } y_i \\ w_{ij}^{*(j)} &= P(y_i = y_i^{*(j)} \mid x_i; \hat{\theta}), \end{aligned}$$

where $\hat{\theta}$ is the (pseudo) MLE of θ.

Fractional Imputation

Parametric fractional imputation

- More generally, we can write $y_i = (y_{i1}, \cdots, y_{ip})$ and y_i can be partitioned into $(y_{i,obs}, y_{i,mis})$.
  1. More than one (say M) imputed values of $y_{mis,i}$: $y_{mis,i}^{*(1)}, \cdots, y_{mis,i}^{*(M)}$ from some (initial) density $h\left(y_{mis,i} \mid y_{obs}\right)$.
  2. Create weighted data set
  $$\left\{\left(w_i w_{ij}^*, y_{ij}^*\right); j = 1, 2, \cdots, M; i = 1, 2 \cdots, n\right\}$$
  where $\sum_{j=1}^{M} w_{ij}^* = 1$, $y_{ij}^* = (y_{obs,i}, y_{mis,i}^{*(j)})$
  $$w_{ij}^* \propto f(y_{ij}^*; \hat{\theta})/h(y_{mis,i}^{*(j)} \mid y_{i,obs}),$$
  $\hat{\theta}$ is the (pseudo) maximum likelihood estimator of θ, and $f\left(y; \theta\right)$ is the joint density of y.
  3. The weight w_{ij}^* are the normalized importance weights and can be called fractional weights.

- EM algorithm by fractional imputation
  1. Initial imputation: generate $y_{mis,i}^{*(j)} \sim h\left(y_{i,mis} \mid y_{i,obs}\right)$.
  2. E-step: compute

$$w_{ij(t)}^{*} \propto f\left(y_{ij}^{*}; \hat{\theta}_{(t)}\right) / h(y_{i,mis}^{*(j)} \mid y_{i,obs})$$

     where $\sum_{j=1}^{M} w_{ij(t)}^{*} = 1$.
  3. M-step: update

$$\hat{\theta}^{(t+1)}: \text{ solution to } \sum_{i=1}^{n} \sum_{j=1}^{M} w_i w_{ij(t)}^{*} S\left(\theta; y_{ij}^{*}\right) = 0,$$

     where $S(\theta; y) = \partial \log f(y; \theta) / \partial \theta$ is the score function of θ.
  4. Repeat Step2 and Step 3 until convergence.
- We may add an optional step that checks if $w_{ij(t)}^{*}$ is too large for some j. In this case, $h(y_{i,mis})$ needs to be changed.

Approximation: Calibration Fractional imputation

- In large scale survey sampling, we prefer to have smaller M.
- Two-step method for fractional imputation:
  1. Create a set of fractionally imputed data with size nM, (say $M = 1000$).
  2. Use an efficient sampling and weighting method to get a final set of fractionally imputed data with size nm, (say $m = 10$).
- Thus, we treat the step-one imputed data as a finite population and the step-two imputed data as a sample. We can use efficient sampling technique (such as systematic sampling or stratification) to get a final imputed data and use calibration technique for fractional weighting.

Approximation: Calibration Fractional imputation

- Step-One data set (of size nM):

$$\left\{ \left(w_{ij}^*, y_{ij}^*\right); j = 1, 2, \cdots, M; i = 1, 2 \cdots, n \right\}$$

and the fractional weights satisfy $\sum_{j=1}^{M} w_{ij}^* = 1$ and

$$\sum_{i \in A} \sum_{j=1}^{M} w_i w_{ij}^* S\left(\hat{\theta}; y_{ij}^*\right) = 0$$

where $\hat{\theta}$ is obtained from the EM algorithm after convergence.

- The final fractionally imputed data set can be written

$$\left\{ \left(\tilde{w}_{ij}^*, \tilde{y}_{ij}^*\right); j = 1, 2, \cdots, m; i = 1, 2 \cdots, n \right\}$$

and the fractional weights satisfy $\sum_{j=1}^{m} \tilde{w}_{ij}^* = 1$ and

$$\sum_{i \in A} \sum_{j=1}^{m} w_i \tilde{w}_{ij}^* S\left(\hat{\theta}; \tilde{y}_{ij}^*\right) = 0$$

Variance estimation for fractional imputation

- Replication-based approach

$$\hat{V}_{rep}(\hat{\eta}_{n,g}) = \sum_{k=1}^{L} c_k \left(\hat{\eta}_{n,g}^{(k)} - \hat{\eta}_{n,g} \right)^2$$

where L is the size of replication, c_k is the k-th replication factor, and $\hat{\eta}_{n,g} = \sum_{i \in A} w_i^{(k)} g(y_i)$ is the k-th replication factor.

Variance estimation for fractional imputation

- For each k, we repeat the PFI method
  1. Generate M imputed values from the same proposal distribution h.
  2. Compute $\hat{\theta}^{(k)}$, the k-th replicate of $\hat{\theta}$ using the EM algorithm in the imputed score equation with replication weight $w_i^{(k)}$.
  3. Using the same imputed values $\tilde{y}_{ij}^*$, the replication fractional weights are constructed to satisfy $\sum_{j=1}^m \tilde{w}_{ij}^{*(k)} = 1$ and

$$\sum_{i \in A} \sum_{j=1}^{m} w_i^{(k)} \tilde{w}_{ij}^{*(k)} S\left(\hat{\theta}^{(k)}; \tilde{y}_{ij}^*\right) = 0$$

Variance estimation for fractional imputation

- Variance estimation of $\hat{\eta}_{FI,g} = \sum_{i \in A} \sum_{j=1}^{m} w_i \tilde{w}_{ij}^{*} g(\tilde{y}_{ij})$ is computed by

$$\hat{V}_{rep}(\hat{\eta}_{FI,g}) = \sum_{k=1}^{L} c_k \left(\hat{\eta}_{FI,g}^{(k)} - \hat{\eta}_{FI,g} \right)^2$$

where

$$\hat{\eta}_{FI,g}^{(k)} = \sum_{i \in A} \sum_{j=1}^{m} w_i^{(k)} \tilde{w}_{ij}^{*(k)} g(\tilde{y}_{ij}).$$

Simulation Study

Simulation 1

- Bivariate data (x_i, y_i) of size $n = 200$ with

$$
\begin{aligned}
x_i &\sim N(3,1) \\
y_i &\sim N(-2 + x_i, 1)
\end{aligned}
$$

- x_i always observed, y_i subject to missingness.
- MCAR ($\delta \sim$ Bernoulli(0.6))
- Parameters of interest
  1. $\theta_1 = E(Y)$
  2. $\theta_2 = Pr(Y < 1)$
- Multiple imputation (MI) and fractional imputation (FI) are applied with $M = 50$.
- For estimation of θ_2, the following method-of-moment estimator is used.

$$
\hat{\theta}_{2,MME} = n^{-1} \sum_{i=1}^{n} I\left(y_i < 1\right)
$$

Simulation Study

Table 1 Monte Carlo bias and variance of the point estimators.

Parameter	Estimator	Bias	Variance	Std Var
θ_1	Complete sample	0.00	0.0100	100
	MI	0.00	0.0134	134
	FI	0.00	0.0133	133
θ_2	Complete sample	0.00	0.00129	100
	MI	0.00	0.00137	106
	FI	0.00	0.00137	106

Table 2 Monte Carlo relative bias of the variance estimator.

Parameter	Imputation	Relative bias (%)
$V(\hat{\theta}_1)$	MI	-0.24
	FI	1.21
$V(\hat{\theta}_2)$	MI	23.08
	FI	2.05

Simulation study

- Rubin's formula is based on the following decomposition:

$$V(\hat{\theta}_{MI}) = V(\hat{\theta}_n) + V(\hat{\theta}_{MI} - \hat{\theta}_n)$$

where $\hat{\theta}_n$ is the complete-sample estimator of θ. Basically, W_M term estimates $V(\hat{\theta}_n)$ and $(1 + M^{-1})B_M$ term estimates $V(\hat{\theta}_{MI} - \hat{\theta}_n)$.

- For general case, we have

$$V(\hat{\theta}_{MI}) = V(\hat{\theta}_n) + V(\hat{\theta}_{MI} - \hat{\theta}_n) + 2Cov(\hat{\theta}_{MI} - \hat{\theta}_n, \hat{\theta}_n)$$

and Rubin's variance estimator ignores the covariance term. Thus, a sufficient condition for the validity of unbiased variance estimator is

$$Cov(\hat{\theta}_{MI} - \hat{\theta}_n, \hat{\theta}_n) = 0.$$

- Meng (1994) called the condition congeniality of $\hat{\theta}_n$.
- Congeniality holds when $\hat{\theta}_n$ is the MLE of θ.

Discussion

- The validity of Rubin's variance formula in MI requires the *congeniality condition* of Meng (1994).
  - Under the congeniality condition:

$$V(\hat{\eta}_{MI}) = V(\hat{\eta}_n) + V(\hat{\eta}_{MI} - \hat{\eta}_n), \tag{1}$$

    where $\hat{\eta}_n$ is the full sample estimator of η. Rubin's formula $\hat{V}_{MI}(\hat{\eta}_{MI}) = W_m + \left(1 + \frac{1}{m}\right) B_m$ is consistent.

  - For general case, we have

$$V(\hat{\theta}_{MI}) = V(\hat{\theta}_n) + V(\hat{\theta}_{MI} - \hat{\theta}_n) + 2Cov(\hat{\theta}_{MI} - \hat{\theta}_n, \hat{\theta}_n) \tag{2}$$

    Rubin's formula can be biased if $Cov(\hat{\theta}_{MI} - \hat{\theta}_n, \hat{\theta}_n) \neq 0$.

  - The congeniality condition holds true for estimating the population mean; however, it does not hold for the method of moments estimator of the proportions.

Discussion

- For example, there are two estimators of $\theta = P(Y < 1)$ when Y follows from $N(\mu, \sigma^2)$.
  1. Maximum likelihood method: $\hat{\theta}_{MLE} = \int_{-\infty}^{1} \phi(z; \hat{\mu}, \hat{\sigma}^2) dz$
  2. Method of moments: $\hat{\theta}_{MME} = n^{-1} \sum_{i=1}^{n} I(y_i < 1)$
- In the simulation setup, the imputed estimator of θ_2 can be expressed as

$$\hat{\theta}_{2,I} = n^{-1} \sum_{i=1}^{n} \left[\delta_i I(y_i < 1) + (1 - \delta_i) E\{ I(y_i < 1) \mid x_i; \hat{\mu}, \hat{\sigma} \} \right].$$

  Thus, imputed estimator of θ_2 "borrows strength" by making use of extra information associated with $f(y \mid x)$.
- Thus, when the congeniality conditions does not hold, the imputed estimator improves the efficiency (due to the imputation model that uses extra information) but the variance estimator does not recognize this improvement.

Simulation 2

- Bivariate data (x_i, y_i) of size $n = 100$ with

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 \left(x_i^2 - 1\right) + e_i \qquad (3)$$

  where $(\beta_0, \beta_1, \beta_2) = (0, 0.9, 0.06)$, $x_i \sim N(0, 1)$, $e_i \sim N(0, 0.16)$, and x_i and e_i are independent. The variable x_i is always observed but the probability that y_i responds is 0.5.
- In MI, the imputer's model is

$$Y_i = \beta_0 + \beta_1 x_i + e_i.$$

  That is, imputer's model uses extra information of $\beta_2 = 0$.
- From the imputed data, we fit model (3) and computed power of a test $H_0 : \beta_2 = 0$ with 0.05 significant level.
- In addition, we also considered the Complete-Case (CC) method that simply uses the complete cases only for the regression analysis

Table 3 Simulation results for the Monte Carlo experiment based on 10,000 Monte Carlo samples.

Method	$E(\hat{\theta})$	$V(\hat{\theta})$	R.B. $(\hat{V})$	Power
MI	0.028	0.00056	1.81	0.044
FI	0.046	0.00146	0.02	0.314
CC	0.060	0.00234	-0.01	0.285

Table 3 shows that MI provides efficient point estimator than CC method but variance estimation is very conservative (more than 100% overestimation). Because of the serious positive bias of MI variance estimator, the statistical power of the test based on MI is actually lower than the CC method.

Summary

- Imputation can be viewed as a Monte Carlo tool for computing the conditional expectation.
- Monte Carlo EM is very popular but the E-step can be computationally heavy.
- Parametric fractional imputation is a useful tool for frequentist imputation.
- Multiple imputation is motivated from a Bayesian framework. The frequentist validity of multiple imputation requires the condition of congeniality.
- Uncongeniality may lead to overestimation of variance which can seriously increase type-2 errors.