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 $\eta_g$
Introduction

Imputation

- 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)
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 \( \eta_g \)

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

Need \( E \{ g(y_i^*) | \delta_i = 0 \} = E \{ g(y_i) | \delta_i = 0 \} \).
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 $\eta_g$ under MAR:

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

Computing the conditional expectation can be a challenging problem.

1. Do not know the true parameter $\theta$ 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 $\theta$, computing the conditional expectation can be numerically difficult.
**Imputation**: Monte Carlo approximation of the conditional expectation (given the observed data).

\[
E \{ g(y_i) \mid x_i \} \approx \frac{1}{M} \sum_{j=1}^{M} g(y_i^{*(j)})
\]

1. **Bayesian approach**: generate \( y_i^* \) from

\[
f(y_i \mid x_i, y_{obs}) = \int f(y_i \mid x_i, \theta) p(\theta \mid x_i, y_{obs}) d\theta
\]

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 $\hat{\eta}_{l,g} = \hat{\eta}_{l,g}(\hat{\theta})$.

To estimate the variance of $\hat{\eta}_{l,g}$, we need to take into account of the sampling variability of $\hat{\theta}$ in $\hat{\eta}_{l,g} = \hat{\theta}_{l,g}(\hat{\theta})$. 
Three approaches

- **Bayesian approach**: multiple imputation by Rubin (1978, 1987), Rubin and Schenker (1986), etc.


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

Idea (parametric model approach)

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

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

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

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

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

$\text{Yang & Kim (ISU)}$ 
$\text{Fractional Imputation}$ 
$\text{October 17, 2014 11 / 28}$
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 $h(y_{mis,i} | y_{obs})$.
2. Create weighted data set

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

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

$$w_{ij}^* \propto f(y_{ij}^*; \hat{\theta}) / h(y_{mis,i}^{*(j)} | y_{i,obs}),$$

$\hat{\theta}$ is the (pseudo) maximum likelihood estimator of $\theta$, and $f(y; \theta)$ is the joint density of $y$.
3. The weight $w_{ij}^*$ are the normalized importance weights and can be called fractional weights.
Proposed method: Fractional imputation

Maximum likelihood estimation using FI

- **EM algorithm** by fractional imputation
  1. **Initial imputation**: generate \( y_{mis, i}^* (j) \sim h(y_{i, mis} \mid y_{i, obs}) \).
  2. **E-step**: compute
     \[
     w_{ij}^*(t) \propto f(y_{ij}^*; \hat{\theta}(t)) / h(y_{i, mis} \mid y_{i, obs})
     \]
     where \( \sum_{j=1}^{M} w_{ij}^*(t) = 1 \).
  3. **M-step**: update
     \( \hat{\theta}(t+1) \) solution to
     \[
     \sum_{i=1}^{n} \sum_{j=1}^{M} w_i w_{ij}^*(t) S(\theta; y_{ij}^*) = 0,
     \]
     where \( S(\theta; y) = \partial \log f(y; \theta)/\partial \theta \) is the score function of \( \theta \).
  4. Repeat Step 2 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.
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 $nM$):
  \[ \left\{ (w_{ij}^*, y_{ij}^*) ; j = 1, 2, \cdots, M; i = 1, 2 \cdots, n \right\} \]
  and the fractional weights satisfy $\sum_{j=1}^{M} w_{ij}^* = 1$ and
  \[ \sum_{i \in A} \sum_{j=1}^{M} w_i w_{ij}^* S (\hat{\theta}; y_{ij}^*) = 0 \]
  where $\hat{\theta}$ is obtained from the EM algorithm after convergence.

- The final fractionally imputed data set can be written
  \[ \left\{ (\tilde{w}_{ij}^*, \tilde{y}_{ij}^*) ; j = 1, 2, \cdots, m; i = 1, 2 \cdots, n \right\} \]
  and the fractional weights satisfy $\sum_{j=1}^{m} \tilde{w}_{ij}^* = 1$ and
  \[ \sum_{i \in A} \sum_{j=1}^{m} w_i \tilde{w}_{ij}^* S (\hat{\theta}; \tilde{y}_{ij}^*) = 0 \]
Replication-based approach

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

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.
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
  
  \[
  \begin{align*}
  x_i & \sim N(3, 1) \\
  y_i & \sim N(-2 + x_i, 1)
  \end{align*}
  \]

- \(x_i\) always observed, \(y_i\) subject to missingness.
- MCAR \((\delta \sim \text{Bernoulli}(0.6))\)
- 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 \(\theta_2\), the following method-of-moment estimator is used.

\[
\hat{\theta}_{2,MME} = n^{-1} \sum_{i=1}^{n} I(y_i < 1)
\]
## Simulation Study

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

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimator</th>
<th>Bias</th>
<th>Variance</th>
<th>Std Var</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_1$</td>
<td>Complete sample</td>
<td>0.00</td>
<td>0.0100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>MI</td>
<td>0.00</td>
<td>0.0134</td>
<td>134</td>
</tr>
<tr>
<td></td>
<td>FI</td>
<td>0.00</td>
<td>0.0133</td>
<td>133</td>
</tr>
<tr>
<td>$\theta_2$</td>
<td>Complete sample</td>
<td>0.00</td>
<td>0.00129</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>MI</td>
<td>0.00</td>
<td>0.00137</td>
<td>106</td>
</tr>
<tr>
<td></td>
<td>FI</td>
<td>0.00</td>
<td>0.00137</td>
<td>106</td>
</tr>
</tbody>
</table>

### Table 2  Monte Carlo relative bias of the variance estimator.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Imputation</th>
<th>Relative bias (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V(\hat{\theta}_1)$</td>
<td>MI</td>
<td>-0.24</td>
</tr>
<tr>
<td></td>
<td>FI</td>
<td>1.21</td>
</tr>
<tr>
<td>$V(\hat{\theta}_2)$</td>
<td>MI</td>
<td>23.08</td>
</tr>
<tr>
<td></td>
<td>FI</td>
<td>2.05</td>
</tr>
</tbody>
</table>
Rubin’s formula is based on the following decomposition:

\[ V(\hat{\theta}_M) = V(\hat{\theta}_n) + V(\hat{\theta}_M - \hat{\theta}_n) \]

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

For general case, we have

\[ V(\hat{\theta}_M) = V(\hat{\theta}_n) + V(\hat{\theta}_M - \hat{\theta}_n) + 2\text{Cov}(\hat{\theta}_M - \hat{\theta}_n, \hat{\theta}_n) \]

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

\[ \text{Cov}(\hat{\theta}_M - \hat{\theta}_n, \hat{\theta}_n) = 0. \]

Meng (1994) called the condition congeniality of \( \hat{\theta}_n \).

Congeniality holds when \( \hat{\theta}_n \) is the MLE of \( \theta \).
The validity of Rubin’s variance formula in MI requires the *congeniality condition* of Meng (1994).

Under the congeniality condition:

\[
V(\hat{\eta}_{MI}) = V(\hat{\eta}_n) + V(\hat{\eta}_{MI} - \hat{\eta}_n),
\]

(1)

where \(\hat{\eta}_n\) is the full sample estimator of \(\eta\). 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

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

(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 $N(\mu, \sigma^2)$.
  - Maximum likelihood method: $\hat{\theta}_{MLE} = \int_{-\infty}^{1} \phi(z; \hat{\mu}, \hat{\sigma}^2) \, dz$
  - Method of moments: $\hat{\theta}_{MME} = n^{-1} \sum_{i=1}^{n} I(y_i < 1)$

- In the simulation setup, the imputed estimator of $\theta_2$ 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 $\theta_2$ “borrows strength” by making use of extra information associated with $f(y \mid x)$.

- Thus, when the congeniality conditions do 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 Study

Simulation 2

- Bivariate data \((x_i, y_i)\) of size \(n = 100\) with
  \[
  Y_i = \beta_0 + \beta_1 x_i + \beta_2 (x_i^2 - 1) + e_i \tag{3}
  \]
  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.

<table>
<thead>
<tr>
<th>Method</th>
<th>$E(\hat{\theta})$</th>
<th>$V(\hat{\theta})$</th>
<th>R.B. ($\hat{V}$)</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>MI</td>
<td>0.028</td>
<td>0.00056</td>
<td>1.81</td>
<td>0.044</td>
</tr>
<tr>
<td>FI</td>
<td>0.046</td>
<td>0.00146</td>
<td>0.02</td>
<td>0.314</td>
</tr>
<tr>
<td>CC</td>
<td>0.060</td>
<td>0.00234</td>
<td>-0.01</td>
<td>0.285</td>
</tr>
</tbody>
</table>

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.
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.